Speech Recognition by the Hearing Impaired

ASHA REPORTS

NUMBER 14 A PUBLICATION OF THE AMERICAN SPEECH-LANGUAGE-HEARING ASSOCIATION

SPEECH RECOGNITION BY THE HEARING IMPAIRED

American Speech-Language-Hearing Association
10801 Rockville Pike • Rockville, Maryland 20852 • (301) 897-5700 (Voice or TTY)

November 1984

Dear Colleague:

The American Speech-Language-Hearing Association is pleased to provide you with a complimentary copy of ASHA Reports 14: Speech Recognition by the Hearing Impaired. This publication is made possible by a grant supported by the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS), National Institutes of Health.

As requested by NINCDS, complimentary copies of ASHA Reports 14 are being sent to all ASHA member audiologists, conference participants, Executive Board members, and Publications Board members. A limited number of copies are available to ASHA members for a $5.00 postage and handling fee.

Our sincere gratitute is expressed to Earleen Elkins, Editor of ASHA Reports 14, and to Theodore J. Glattke, ASHA Reports Series Editor, as well as to the following authors for their contributions:

Robert C. Bilger
Louis D. Braida
Marilyn E. Demorest
Earleen Elkins
John J. Godfrey
Deborah Hayes
Candace A. Kamm
Harry Levitt
James D. Miller
Steffi B. Resnick
Charles E. Speaks
Aaron R. Thornton
Brian E. Walden

Sincerely,

Frederick T. Spahr
Executive Director

SPEECH RECOGNITION BY THE HEARING IMPAIRED

Edited by

EARLEEN ELKINS, PH.D.

National Institute of Neurological and
Communicative Disorders and Stroke
National Institutes of Health

ASHA Reports 14 (ISSN 0569-8553)
AMERICAN SPEECH-LANGUAGE-HEARING ASSOCIATION
Rockville, Maryland
October 1984

AMERICAN SPEECH-LANGUAGE-HEARING ASSOCIATION

President
Chair, Legislative Council
Chair, Executive Board

David E. Yoder, Ph.D.
University of Wisconsin—Madison

EXECUTIVE BOARD

President-Elect
Richard M. Flower, Ph.D.
University of California, San Francisco

Past President
Fred D. Minifie, Ph.D.
University of Washington, Seattle

Vice President for Administration
Sandra C. Holley, Ph.D.
Southern Connecticut State University
New Haven

Vice President for Clinical Affairs
Hughlett L. Morris, Ph.D.
University of Iowa, Iowa City

Vice President for Education and Scientific Affairs
Steffi B. Resnick, Ph.D.
J. F. Kennedy Institute
Baltimore, Maryland

Vice President for Planning
Carol P. Leslie, Ph.D.
Cincinnati Speech and Hearing Center
Cincinnati, Ohio

Vice President for Professional and Governmental Affairs
Mariana Newton, Ph.D.
University of North Carolina at Greensboro

Vice President for Standards and Ethics
Jane R. Madell, Ph.D.
New York League for the Hard of Hearing
New York, New York

Executive Director
Frederick T. Spahr, Ph.D.

PUBLICATIONS BOARD

Robert C. Bilger, Ph.D., **Chair**
Patricia A. Broen, Ph.D.
Robin S. Chapman, Ph.D.
Janis M. Costello, Ph.D.
Theodore J. Glattke, Ph.D.
Kathleen M. Griffin, Ph.D. (*ex officio*)
Miriam A. Henoch, Ph.D.
John L. Locke, Ph.D.
Arnold M. Small, Ph.D.
Lynn S. Snyder, Ph.D.
Frederick T. Spahr, Ph.D.
C. W. Starkweather, Ph.D.

ASHA REPORTS NUMBER 14

Editor

Earleen Elkins, Ph.D.

Series Editor
Theodore J. Glattke, Ph.D., 1979–83
Arnold M. Small, Ph.D., 1984–

Business Manager
Frederick T. Spahr, Ph.D.

Publications/Marketing Division
Alfred K. Kawana, Director

Production Editor
Donna Manuel Aspinall

CONTENTS

ACKNOWLEDGMENTS

Publication of this report is supported by the National Institute of Neurological and Communicative Disorders and Stroke. Special thanks are expressed to Jerry Raher who provided secretarial management with extraordinary skill.

CONFERENCE PARTICIPANTS

Robert C. Bilger, Ph.D.	University of Illinois, Champaign, Illinois
Louis D. Braida, Ph.D.	Massachusetts Institute of Technology, Cambridge, Massachusetts
Marilyn E. Demorest, Ph.D.	University of Maryland Baltimore County, Baltimore, Maryland
Earleen Elkins, Ph.D.	Communicative Disorders Program, National Institute of Neurological and Communicative Disorders and Stroke, National Institutes of Health, Bethesda, Maryland
John J. Godfrey, Ph.D.	University of Texas at Dallas, Dallas, Texas
Deborah Hayes, Ph.D.	Baylor College of Medicine, Houston, Texas
Candace A. Kamm, Ph.D.	American Telephone and Telegraph Information Systems Laboratory, Lincroft, New Jersey
Harry Levitt, Ph.D.	City University of New York, New York, New York
James D. Miller, Ph.D.	Central Institute for the Deaf, St. Louis, Missouri
Steffi B. Resnick, Ph.D.	John F. Kennedy Institute for Handicapped Children, Baltimore, Maryland
Charles E. Speaks, Ph.D.	University of Minnesota, Minneapolis, Minnesota
Aaron R. Thornton, Ph.D.	Massachusetts Eye and Ear Infirmary, Boston, Massachusetts
Brian E. Walden, Ph.D.	Walter Reed Army Medical Center, Washington, DC

Chapter 1

INTRODUCTION

BRIAN E. WALDEN

Walter Reed Army Medical Center
Washington, DC

In late 1982, a group of 12 clinicians and researchers met for 2 days to discuss speech recognition by the hearing impaired. This hardly qualifies as a new topic, and certainly this is not the first report devoted to the subject. The study and assessment of speech recognition have been fundamental to the disciplines of audiology and hearing science since their beginnings. The extensive clinical and experimental attention given to this topic reflects the fact that the recognition of speech, more than any other class of sounds, is critical to effective daily living. It is hardly necessary to say that speech recognition ability is basic to communication. Most of the clinical efforts of rehabilitative audiologists are directed toward restoring this one ability.

Given the long history of clinical and experimental interest in speech recognition by the hearing impaired, one might reasonably ask, "Why this report?" and "Why now?" These questions may best be answered by considering what has been done and what needs to be done in the study of speech recognition by the hearing impaired.

Historically, syllable-length test lists (usually containing 50 items) presented at a single intensity and scored as percent correct have dominated the assessment of suprathreshold speech recognition ability in the hearing impaired. Such an approach was almost universally accepted, particularly in clinical practice, for decades. During the past several years, however, this approach has been increasingly criticized from a variety of perspectives. The diversity of the criticism reflects both the simplicity of the traditional approach and the complexity of speech recognition by the hearing impaired.

Clearly, both clinicians and experimentalists feel the need to move in new directions—and many of them are. In recent years, several new test materials, methodologies, and analyses have been suggested. Yet, due to a variety of limitations, none have received widespread acceptance. As a result, despite our substantially increased knowledge of speech recognition by the hearing impaired, old testing approaches largely persist. Although this may be disturbing, it is probably quite predictable. In the absence of a universally acceptable alternative, and faced with the necessity of doing something, clinicians and researchers are likely to resort to old, familiar approaches despite their obvious limitations. Clearly, much more basic and clinical research into speech recognition by the hearing impaired is required before these problems will be resolved.

This report is an effort to add direction to such future research efforts. Its purpose is to study speech recognition by the hearing impaired from a broad perspective. Discussions will include psychometric, linguistic, acoustic, and methodological considerations that are not restricted to or by a single investigator's data and experimental approaches. This report is not intended to be simply a research report by a dozen of the leading experts on speech recognition by the hearing impaired.

At the end of the report, a summary of future research directions in speech recognition by the hearing impaired is presented. As is true of most groups of clinicians and researchers, all of them did not agree on each suggestion but there was unanimity in the concern that speech recognition by the hearing impaired is still an area ripe for meaningful study.

Chapter 2

SPEECH RECOGNITION TEST DEVELOPMENT

ROBERT C. BILGER

University of Illinois, Champaign

Although tests of speech recognition may be considered to be tests of sensory capacity, in their development and standardization they are completely analogous to tests of mental ability. Thus, the principles of psychometric theory and the conventional wisdom of psychometric practice should apply to the development and standardization of tests of speech recognition. From the standpoint of psychometric practice (Linquist, 1953), it is possible to specify seven steps in the process of developing and standardizing a test of speech recognition:

1. Define the test.
2. Prepare a large pool of prospective test items.
3. Pretest all this pool of items by administering them to a large number of subjects drawn from the population to which the final form of the test will be administered.
4. Conduct a psychometric evaluation of the prospective test items using data from part of the sample.
5. Construct one or more forms of the test to meet all criteria.
6. Cross-validate the test using data from those subjects held out of the initial psychometric evaluation.
7. Validate the final forms of the test on a new sample of appropriate subjects.

If the seventh step does not indicate that the test meets all the criteria set for it, then it will be necessary to return to an earlier step and repeat the sequence again. It will be convenient here to integrate the requisite theoretical principles into an elaboration of the seven steps.

Definition of the Test

The first step in developing a test requires a series of stages through which the developers must pass. The first stage is to specify the construct that the test is to measure. The second stage relates to the kind of test items that are to be used in measuring the specified construct and involves determining the domain of possible test items to be used, the procedure for sampling that domain and the exact format of the test items. At a third stage in defining the test, one must consider and select the kind of response to be used and must determine how responses are to be scored. In a discussion of the relevance of these stages in the definition of a test, a major emphasis must be placed on the development of the concept of a construct.

Identification of the construct to be measured by a test is of primary importance both from a practical and a theoretical point of view. From a practical point of view, the identification of the construct is essential to assessing the validity of a test and to specifying the applications of the test. However, the major importance of identifying the construct is theoretical (Nunnally, 1978). In this context, the dictionary defines the word *construct* as "something constructed, especially by mental synthesis; as the concept of a physical object from sense data." The variables to be measured are distributed along a continuum from concrete to abstract, and abstract variables deal with constructs. Abstract variables or constructs are usually generated by scientific theory in the scientist's imagination or by the reality of clinical practice. Concern about the need for and the difficulty in specifying the validity and reliability of a test is a good indication that one is dealing with a construct. In practice, constructs are inferred from a domain of observables, and the more systematically the domain of observables is studied, the better the construct will be established. Ultimately, constructs involve inferences that a scientist wants to draw from an experiment or that a clinician wants to draw from the results of the tests administered to an individual.

In fairness to the developers of tests of speech recognition, it should be noted that concern with constructs measured by tests (Cronbach & Meehl, 1955) arose after the initial application of word lists to the measurement of speech recognition (Egan, 1948; Hudgins, Hawkins, Karlin, & Stevens, 1947). Though the original phonetically balanced word lists were developed to provide an experimental technique for evaluating radio/telephone systems, that development was based on a sound construct of obtaining a reasonable sample of the spectrum of speech from each of the experimental talkers. In Egan's investigation, the speech of a sample of talkers was transmitted through a radio/telephone system (the independent variable) to a panel of listeners, whose mean score provided the criterion (the dependent variable) against which systems were evaluated. Adoption of these word lists for testing speech recognition in individual listeners brought with it several problems that are related to the second stage of defining a test of speech recognition. These include identifying the domain of possible items, sampling that domain, and selecting a format for the items. In the radio/telephone experiment, the orthographic representations of the words were the items of the test. The validity of the items was evaluated in terms of how effectively the talker generated a reasonable representation of the spectrum of speech. As a measure of speech recognition, however, the test item is really an acoustic event, the utterance of the word; and the validity of these utterances depends on how effectively they differentiate among listeners with good speech recognition and those with poor speech recognition.

Thus, one characteristic of the domain of test items to be sampled in constructing a test of speech recognition is that it involves spoken words, not printed words. Although the first recording of monosyllabic words to be used as a test of speech recognition used a typical talker (Hudgins et al., 1947), most available recordings have used talkers whose articulatory gestures are atypically precise. The excellence of the utterances of these talkers, coupled with the fact that words uttered in isolation tend to be more readily understood than the individual words in connected discourse, result in estimates of speech recognition ability that most likely underestimate an individual's difficulty in understanding speech. Perhaps the domain of items to be sampled should be defined to cover the range of quality of the utterances, as well as the traditional dimensions of phonemic content.

In conventional tests, the individual's response is dependent upon the format of the items: "Say the word, *cheese*, please" or "Number 33 is *cowl.*" Although writing down or repeating the test word is convenient, such parroting is relatively atypical of the manner in which people respond to speech in real-life situations. The classic format-response problem in speech recognition testing involves the CHABA[1] Sentence Test (CID Sentences—Giolas & Duffy, 1973). In this case, only after the sentences were written and recorded did concern about the response to the sentences and the scoring of those responses arise. The keyword scoring scheme is generally unsatisfactory (Giolas & Duffy, 1973). Any developer of speech recognition tests should devote some thought to the construction of items that are realistic samples of speech, for which the response to and scoring of the items is inherent in the items themselves and for which the response is more realistic than the conventional repeat-back method.

Examples of the desired creativity in test construction can be found in the Speech Perception in Noise (SPIN) Test (Kalikow, Stevens, & Elliott, 1977) and in the Token Test (DeRenzi & Vignollo, 1962). The SPIN Test provides an excellent example of test items that approximate real-life speech, even though the individual's response is only to repeat the last word of the sentence. The Token Test, designed to test language comprehension of individuals with aphasia, uses rather stilted speech but the responses are simple and appropriate nonspeech, motor actions by the subject.

Preparation, Pretest, and Evaluation of Prospective Test Items

When all aspects of the test have been defined, the second step is to prepare a large pool of test items. Construction of good test items is a difficult task, but obviously poor items may be eliminated by having several judges read the material. After the materials have been recorded, the items must be pretested by presenting them to a relatively large sample of subjects drawn from the population of people to whom the test is to be applied. The goal of such pretesting of potential items is to determine the test items that effectively discriminate between listeners with good and those with poor speech recognition. The correlations being sought are called *item validities* (Nunnally, 1978).

Test theory assumes that the score obtained on a particular administration of a test has two components, the individual's *true score* and an *error score*. This can be expressed as

$$X_i = T_i + E \tag{1}$$

where X represents the *obtained score*, T the *true score*, and E the *error score*. One can demonstrate that

$$\sigma_x{}^2 = \sigma_t{}^2 + \sigma_E{}^2 \tag{2}$$

by assuming that true scores and error scores are uncorrelated and that the mean error score for the population is zero. Following Nunnally (1978), reliability can be defined as the ratio of *true variance* to *obtained variance*.

$$r_{xx} = \frac{\sigma_t{}^2}{\sigma_x{}^2} = \frac{\sigma_t{}^2}{\sigma_t{}^2 + \sigma_E{}^2}\,. \tag{3}$$

The concepts of *true score* and *true variance* require explanation. The concept of *true score* assumes that similar test items are drawn from a population or domain of such items so that an individual's true score is the score he/she would obtain if tested on all of the items in the domain. The correlation between true score and scores on a particular form of a test, when squared, becomes an alternative definition of reliability. Further, the correlation between *item scores* and *true scores* represents the validity of the test item as a measure of that domain.

The operational definition of *true variance* is extremely relevant here. Given the definition of reliability in Equation 3, one can recognize that the various paradigms for assessing the reliability of a test (split-halves, test-retest, or equivalent forms) are procedures for estimating the terms in Equation 3. If the split-halves paradigm is extended to its limit and includes the interitem correlations, then it becomes convenient to use a random-model analysis of variance to estimate the three variance terms in Equation 2. Using this approach, reliability can be defined (Winer, 1971) as

$$r_{xx} = \frac{MS_{\text{subj}} - MS_{\text{error}}}{MS_{\text{subj}} + (k-1)\,MS_{\text{error}}}\,. \tag{4}$$

Equation 4 infers that the *true variance* of test theory is estimated by the *among-subjects variance* component in the random-model analysis of variance. This among-subjects variance (the variance among subjects less the error variance divided by the number of test items on which the subjects' totals are based) is an unbiased estimate of individual differences. One can then infer that the prime requisite of a reliable test is that it measure a trait for which individual differences are extremely large with respect to the error of measurement. Since reliability can be defined in terms of true variance, which is best obtained from the unbiased esti-

[1] Committee on Hearing and Bioacoustics and Biomechanics.

mate of among-subjects variance calculated from a random-model analysis of variance, it seems apparent that all phases of test development should use representative samples drawn from the population to which the test is to be applied.

Many clinicians express concern over the need to use a sample chosen to estimate the population value of *among-subjects variance.* This concern probably relates to the question of whether error of measurement for a specified test is independent of degree, etiology, or age at the onset of hearing loss. Such concerns may be resolved by recognizing that questions about the reliability of a test, the equivalence of forms of a test, or the validity of a test ultimately involve the need to estimate the magnitude of among-subjects variance in the population for which the proposed test is being developed and that error of measurement is statistically independent of among-subjects variance. If the proposed test is reliable, then the error of measurement is a very small component of the denominator in Equation 3. Although the question about the independence of error variance from clinical category is an important clinical concern, it does not impact significantly on issues of reliability (equivalence) or validity.

Construction of Test Forms and Cross-Validation

Following the tape-recorded presentation of the proposed test of speech recognition to a large number of subjects from the population of interest (hearing-impaired subjects in this case), the effectiveness or validity of each item should be determined. If the original pool of prospective test items was large enough, then the subjects' total scores on all the items can be used as an estimate of true score (Nunnally, 1978). In this case, the correlation between the score on each item and true score will be the validity of the item and the square of that correlation will be the reliability of that item. The wise test developer will use a hold-out technique in calculating these correlations, so that part of this initial sample can be used to cross-validate the test forms to be constructed on the basis of the item difficulties and the item reliabilities. That is, given that the item reliabilities and validities are known, it is now practical to construct several forms of the proposed test of speech recognition that will meet the psychometric criteria for equivalence (equal means, equal variances, and equal correlations with true score) and also audiological criteria.

For example, Bilger, Neutzel, Rabinowitz, and Rzeczkowski (1984) administered the SPIN Test (Kalikow et al., 1977) to 128 subjects with sensorineural hearing loss and found that the 10 forms were not equivalent with respect to any of the three psychometric criterion of equal difficulty, homogeneity of variance, or equal correlations between obtained and true scores. The mean scores, standard deviations, and reliability coefficients are summarized in Table 1 for the high- and low-predictability subtests of the SPIN Test. The lack of equivalence is not surprising when one considers that equivalent forms were constructed on the basis of pretesting tape-recorded items and the talker later reuttered the equivalent forms. Obviously, the new utterances were a different test.

TABLE 1. Psychometric data from standardization of original SPIN Test (Bilger et al., 1984). Calculations are based on $N = 128$. Reliability coefficients are the mean of interform correlations.

	High-predictability items			*Low-predictability items*		
Form #	*Mean*	*SD*	r_{xx}	*Mean*	*SD*	r_{xx}
1	89.11	17.30	.917	45.08	21.41	.818
2	88.64	17.00	.904	45.16	22.28	.845
3	87.28	20.07	.919	42.00	22.23	.853
4	84.58	22.58	.919	52.09	22.09	.863
5	89.45	18.90	.910	50.69	23.89	.866
6	89.44	18.02	.922	54.80	25.44	.846
7	83.80	22.53	.918	42.86	21.62	.832
8	85.55	20.73	.900	36.30	21.29	.850
9	85.56	20.03	.894	41.81	23.26	.858
10	87.88	19.14	.915	47.62	23.74	.855
$\bar{x}$	87.13	19.82	.905	45.84	23.36	.850

To determine the constituency of equivalent forms, the data derived from the first 64 subjects of the revised version (Bilger et al., 1984) were analyzed to determine each test item's difficulty and correlation with the total scores summed across all 250 low- and all 250 high-predictability test items of the original SPIN Test. That is, these total scores were used as true scores. Based on this item analysis, 14 words failed to correlate significantly with their true score (Table 2). From the remaining pool of 236 words, four sets of 50 words each were selected so that the mean and variance of item difficulties and item validities were essentially equal. Once equality had been achieved for difficulty, variance, and reliability, words of equal difficulty and item validity were exchanged from list to list to achieve balance with respect to syllable, vowel, and consonant type. The results of this process are summarized in Tables 3 and 4, and the final forms of the Revised SPIN Test are presented as an Appendix. The psychometric data are summarized in Table 3. In this table, the results for the test sample of 64 subjects and the hold-out sample are displayed separately for high- and low-predictability items. The results, in terms of phonetic parameters, are summarized in Table 4.

Although the results in Table 3 indicate that a high degree of equivalence could be obtained, the Revised SPIN Test did not exist yet. To this end, the recordings of the SPIN Test were digitized and stored on disk files of a large minicomputer. The digitization of the SPIN Test was run at a rate of 40,000 samples/s. The analog-to-digital converter was multiplexed so that alternate samples were taken from the speech and babble tracks of the tapes, giving an effective rate of 20,000 Hz for each. For playback digital-to-analog conversion, the speech and noise were low-pass filtered at 7000 Hz to eliminate aliases from the output.

Validation of Final Test Forms

The tapes of the Revised SPIN Test were then played to 32 subjects who had sensorineural hearing loss. These 32 subjects were selected from the original 128 subjects. The results of the retesting with the Revised SPIN are summarized in Table 5. These data indicate that the difficulty and variance of the eight forms were still present in the actual Revised SPIN Test, but that the reliability of Form 2 of the Revised SPIN

TABLE 2. Difficulty and item validity of words from the SPIN Test that were deleted from the Revised SPIN Test for low- and high-predictability contexts. Difficulty is the proportion of correct responses for 64 subjects in test sample. Item validity is correlation of item scores with total score for low- or high-predictability items.

	Low-predictability context				*High-predictability context*			
		Criterion			*Criterion*			
Form-item #	*Difficulty*	*High*	*Low*	*Test word*	*High*	*Low*	*Difficulty*	*Form-item #*
1-11	.219	.228*	.290	hive	.136*	−.078*	.812	2-26
2-38	.234	−.014*	.375	booth	.244*	.244*	.969	1-32
1-20	.016	.058*	.071*	hug	.406	.342	.844	2-34
1-36	.375	.098*	.040	lock	.423	.291	.859	2-40
3-6	.094	.131*	.221*	dusk	.810	.634	.859	4-38
7-48	.091	.091*	.144	cramp	.544	.582	.672	8-27
8-45	.094	.150*	.107*	curb	.603	.428	.828	7-8
10-28	.297	.175*	.189*	beak	.368	.289	.719	9-2
7-12	.500	.181*	.204*	plot	.527	.522	.734	8-13
8-25	.109	.144*	.169*	chat	.838	.545	.922	7-14
8-46	.016	.077*	.173*	tin	.682	.532	.812	7-46
9-13	.438	.134*	.219*	vault	.788	.496	.937	10-17
2-27	.312	.396	.148*	draft	.747	.493	.938	1-9
4-13	.516	.303	.244*	curve	.586	.468	.891	3-3
1-25	.094	.147*	.260	pine	.406	.342	.844	2-47
1-40	.438	.123*	.261	track	.640	.421	.906	2-5
2-15	.141	.198*	.295	scab	.562	.523	.797	1-35
3-4	.156	.081*	.252	band	.747	.493	.983	4-19
4-12	.094	.183*	.281	bud	.633	.578	.766	3-35
5-19	.297	.039*	.303	limb	.426	.314	.938	6-20
6-5	.172	.200*	.258	robe	.713	.490	.922	5-11
7-20	.203	.236*	.338	mop	.438	.327	.938	8-50
7-39	.703	.230*	.283	meal	.648	.523	.797	8-2
10-15	.141	.196*	.312	beam	.628	.478	.906	9-38
1-30	.172	.227*	.401	pad	.572	.569	.672	2-8
9-29	.156	.217*	.330	cop(s)	.752	.562	.844	10-47
1-33	.656	.444	.373	cake	.689	.464	.938	2-4
2-9	.625	.304	.578	bloom	.579	.337	.969	1-18
2-24	.297	.271	.321	crop	.508	.548	.672	1-17
3-28	.812	.550	.411	dock	.841	.551	.922	4-4
4-23	.547	.329	.395	peak	.461	.274	.938	3-49
4-29	.500	.315	.278	rake	.698	.515	.906	3-46
4-32	.875	.742	.593	bill	.659	.367	.969	3-30
5-40	.422	.315	.361	cot	.617	.448	.906	6-2
6-45	.297	.255	.306	goal	.478	.325	.891	5-36
6-50	.453	.322	.617	blast	.579	.337	.969	5-17
5-37	.922	.352	.267	blame	.516	.285	.984	6-7
9-47	.344	.327	.482	glue	.716	.428	.953	10-44
1-23	.797	.655	.501	doll	.548	.412	.906	2-44
2-2	.391	.387	.295	swamp	.462	.300	.906	1-15
5-9	.516	.403	.390	barn	.716	.428	.953	6-18
6-27	.328	.321	.407	drum	.328	.358	.781	5-24
7-9	.438	.324	.399	bow	.747	.493	.938	8-18
7-27	.297	.311	.442	sling	.614	.482	.875	8-41
8-29	.438	.428	.602	film	.559	.567	.734	7-26
9-30	.531	.408	.438	debt	.516	.258	.984	10-14
9-31	.172	.262	.485	dove	.580	.543	.797	10-1
9-44	.703	.310	.373	flock	.474	.422	.828	10-21
9-46	.219	.315	.374	crumbs	.532	.443	.703	10-35
10-22	.266	.316	.478	tent	.425	.393	.812	9-15

*Correlation failed to reach statistical significance.

Test was slightly lower than the reliability of the other seven forms. Perhaps the user of the Revised SPIN Test should avoid using Form 2 and its cognate, Form 1, where possible.

The foregoing discussion has dealt with the SPIN and Revised SPIN Tests as two separate tests composed of high-predictability items and low-predictability items. Kalikow et al. (1977) recommended using the difference between scores on the two subtests as an indicator of how patients use context. In the analysis of the data on the 128 subjects with sensorineural hearing loss tested in the Bilger et al. study (1984), there was no evidence of differences among subjects regarding the use of context. For that reason, the reliability of the difference score is extremely low, as the difference score between two highly reliable subtests must be.

An alternate strategy for combining the scores of two SPIN subtests would be to add them together. Although the reliability of the high- and low-predictability subtests are each highly reliable, it is inappropriate to use this total score as a

TABLE 3. Psychometric data for eight equivalent forms of the Revised SPIN Test. Reliability coefficients are based on average item-validity and the Spearman-Brown Prophecy Formula.

	Test sample (n = 64)			*Holdout group (n = 64)*		
Form #	*Mean*	*SD*	*rel*	*Mean*	*SD*	*rel*
			High-predictability items			
1	88.64	18.60	.926	86.64	22.80	.937
2	87.36	19.92	.930	84.96	20.92	.926
3	87.52	18.12	.915	85.36	20.72	.932
4	88.80	19.24	.938	85.32	21.00	.930
5	82.24	19.00	.934	85.76	21.44	.943
6	88.68	17.76	.932	85.84	21.36	.949
7	88.16	19.64	.933	85.84	22.32	.935
8	90.64	17.68	.915	88.44	21.20	.937
			Low-predictability items			
1	49.52	25.64	.903	43.32	21.84	.877
2	52.16	24.52	.906	46.24	21.92	.868
3	49.76	24.60	.901	46.00	22.56	.864
4	50.24	24.64	.902	47.20	22.64	.885
5	50.24	25.32	.895	47.80	23.24	.876
6	49.20	26.32	.896	43.80	22.12	.871
7	49.52	24.92	.907	43.88	21.24	.871
8	50.40	25.56	.891	44.08	23.04	.862

TABLE 4. Distribution of (A) vowels, (B) consonants, and (C) syllable type in Revised SPIN lists.

	RO1	*RO2*	*RO3*	*RO4*	*RO5*	*RO6*	*RO7*	*RO8*
A. Vowel type								
High front	5	8	6	4	4	5	6	5
Low front	7	6	7	6	7	7	5	5
Low back	5	3	4	6	6	5	5	5
High back	1	1	1	2	2	1	2	2
Diphthong	2	2	2	2	2	3	2	3
Diphthong +	5	5	5	5	4	4	5	5
B. Consonant type								
Voiceless plosive	16	16	19	15	16	15	18	21
Voiced plosive	9	11	7	6	11	12	8	8
Voiceless fricative	11	10	13	16	14	12	15	14
Voiced fricative	4	4	4	6	3	4	4	4
Semivowel	10	13	12	11	14	13	14	13
Nasal	9	9	9	8	6	8	8	5
Total consonants	59	63	64	62	64	64	67	65
C. Syllable type +[a]								
CN/NC	2	2	1	2	1	2	1	2
CCN/NCC	1	1	1	1	1	1	1	1
CNC	9	10	10	11	9	9	7	9
CNCC	6	5	7	5	6	6	7	6
CCNC	6	6	5	4	7	6	8	6
CCNCC	1	1	1	2	1	1	1	1

[a] C = Consonant; N = Vowel nucleus; CC = Consonant cluster (2 or 3 consonants).

TABLE 5. Validation study of the Revised SPIN Test based on retesting of 32 subjects. Reliabilities were obtained by correlating scores on each form with the subjects' total scores.

	High-predictability items			*Low-predictability items*		
Form #	*Diff*	*Var*	*Rel*	*Diff*	*Var*	*Rel*
1	77.84	22.50	.917	34.05	25.00	.912
2	77.30	25.67	.938	39.46	22.00	.867
3	72.43	25.50	.927	37.30	26.17	.917
4	77.84	25.00	.941	35.14	24.00	.898
5	74.05	23.83	.919	37.30	25.17	.909
6	77.30	25.67	.939	36.76	27.33	.928
7	74.05	25.83	.933	41.62	28.00	.925
8	77.30	25.73	.942	34.05	27.67	.931

clinical measure of speech recognition because its coefficient of generalizability (Cronback, Gleser, Nanda, & Rajaratnam, 1972) is very low—0.38 on the original SPIN.

Any optimum scoring system for combining the score from the high- and low-predictability subtests should be based on the relation between the subtest scores. This relation is shown in Figure 1 for the 128 subjects (Bilger et al., 1984), which shows the high-predictability score on all 10 forms of the original SPIN Test as a function of the corresponding low-predictability score (on the abscissa). Each point is based on 250 words in each context.

Figure 1 has two interesting features. First, because the data are so tightly though not linearly related to one another, they suggest that the two subtests are measuring the same construct—speech recognition. Were one to plot the relation between other measures of speech recognition in a similar manner, one would find comparable relations of spondaic words to monosyllabic words and among nonsense syllable, two- and three-syllable words (Bilger, Nuetzel, Trahiotis, & Rabinowitz, 1980). It is important to note that the strong relation of high- to low-predictability scores is curvilinear, so a Pearson product-moment correlation used to estimate concurrent validity would seriously underestimate the degree of relationship between any two of these speech measures. The fact that the two SPIN subtests and other measures of speech recognition are so closely related suggests that speech recognition is a viable and valid construct.

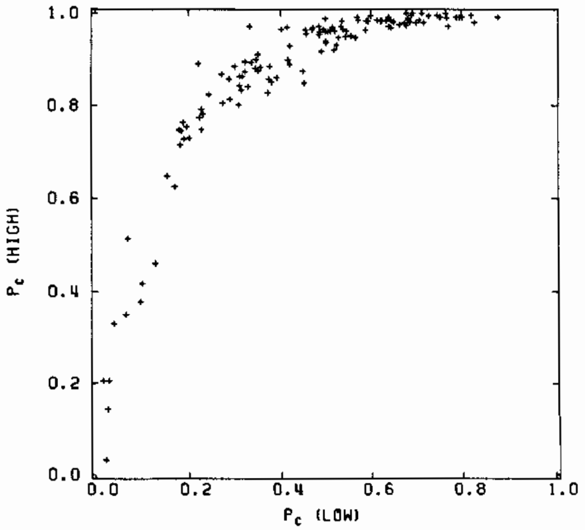

FIGURE 1. Proportion of correct responses to high-predictability items as a function of proportion of correct responses to low-predictability items of the SPIN Test. Each datum point (+) is based on 250 items of high- and low-predictability context.

Second, the relationship seen in Figure 1 suggests that the difference between high- and low-predictability scores is essentially the same for all subjects and that a useful scoring system is possible. To illustrate both of these points, the data were normalized by a normal-transform and replotted in normal-normal space, as shown in Figure 2. The use of a normal transformation here does not require one to assume that the underlying traits are normally distributed and was used only because it is a convenient exponential form.

Excluding data for which the scores on low-predictability items were below 2% or the scores on high-predictability items were above 98%, the line fitted to the data in Figure 2 has a slope of .998. The relevant aspect of a patient's performance is his or her position on this line. The position can be calculated from the average of the *z* scores (normal variate) for the high- and low-predictability subtests. A nomograph for scoring the Revised SPIN Test is available (Bilger, 1984).

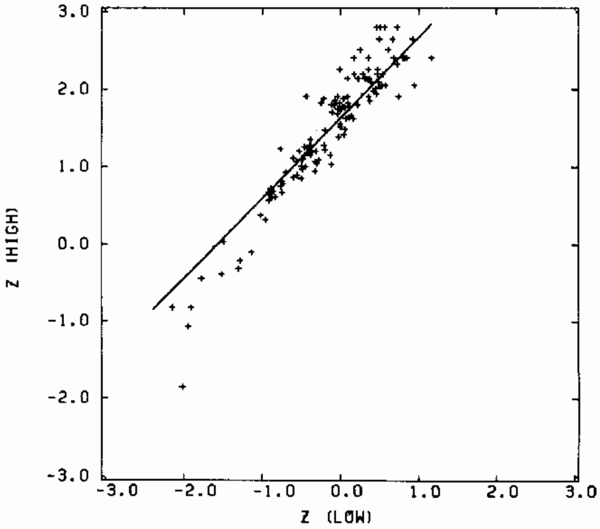

FIGURE 2. Data from Figure 1 have been transformed and replotted on normal-normal coordinates. Data corresponding to proportions greater than 0.98 for high- or less than 0.02 for low-predictability items were excluded from process of fitting the line to these data points.

In summary, seven steps in the development of standardization of tests of speech reception have been discussed or illustrated. The first step, definition of the test, is the most complicated because it involves identifying the construct to be measured, specifying the nonpsychometric criteria for the test, and defining the test items and responses. The second through seventh steps, on the other hand, are routine to psychometrics. These last six steps were illustrated with data concerning the SPIN Test. Because the SPIN Test involves two subtests (high- and low-predictability items), the illustration also afforded an example of construct validity.

REFERENCES

BILGER, R. C. (1984). *Manual for the clinical use of the Revised SPIN Test*. Champaign–Urbana: University of Illinois.

BILGER, R. C., NUETZEL, J. M., RABINOWITZ, W. M., & RZECZKOWSKI, C. (1984). Standardization of a test of Speech Perception in Noise. *Journal of Speech and Hearing Research, 27*, 32–48.

BILGER, R. C., NUETZEL, J. M., TRAHIOTIS, C., & RABINOWITZ, W. M. (1980). An objective psychophysical approach to measuring hearing for speech. *Asha, 22*, 726.

CRONBACH, L. J., GLESER, G. C., NANDA, H., & RAJARATNAM, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

CRONBACH, L. J., & MEEHL, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

DERENZI, E., & VIGNOLLO, L. A. (1962). The Token Test: A sensitive test to detect receptive disturbance in aphasics. *Brain, 85*, 665–678.

EGAN, J.P. (1948). Articulation test methods. *Laryngoscope, 58*, 955–991.

GIOLAS, T. G., & DUFFY, J. (1973). Equivalency of CID and revised CID sentence lists. *Journal of Speech and Hearing Research, 16*, 739–743.

HUDGINS, C., HAWKINS, J., KARLIN, J., & STEVENS, S. (1947). The development of recorded auditory tests for measuring hearing loss of speech. *Laryngoscope, 57*, 57–89.

KALIKOW, D. N., STEVENS, K. N., & ELLIOTT, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61*, 1337–1351.

LINQUIST, E. (1953). *Educational measurement*. Washington, DC: American Council on Education.

NUNNALLY, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

WINER, B. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

APPENDIX

Form #1 of the Revised SPIN Test

Name ________________ (# ________) Marker ________ Date ________
S/B +8 dB #C-HIGH ________ #C-LOW ________ ACCEPT? Y / N Percent Hrg. ________

1. His plans meant taking a big *RISK*. H 1.________
2. Stir your coffee with a *SPOON*. H 2.________
3. Miss White won't think about the *CRACK*. L 3.________
4. He would think about the *RAG*. L 4.________
5. The plow was pulled by an *OX*. H 5.________
6. The old train was powered by *STEAM*. H 6.________
7. The old man talked about the *LUNGS*. L 7.________
8. I was considering the *CROOK*. L 8.________
9. Let's decide by tossing a *COIN*. H 9.________
10. The doctor prescribed the *DRUG*. H 10.________

11. Bill might discuss the *FOAM*. L 11.________
12. Nancy didn't discuss the *SKIRT*. L 12.________
13. Hold the baby on your LAP. H 13.________
14. Bob has discussed the *SPLASH*. L 14.________
15. The dog chewed on a *BONE*. H 15.________
16. Ruth hopes he heard about the *HIPS*. L 16.________
17. The war was fought with armored *TANKS*. H 17.________
18. She wants to talk about the *CREW*. L 18.________
19. They had a problem with the *CLIFF*. L 19.________
20. They drank a whole bottle of *GIN*. H 20.________

21. You heard Jane called about the *VAN*. L 21.________
22. The witness took a solemn *OATH*. H 22.________
23. We could consider the *FEAST*. L 23.________
24. Bill heard we asked about the *HOST*. L 24.________
25. They tracked the lion to his *DEN*. H 25.________
26. The cow gave birth to a *CALF*. H 26.________
27. I had not thought about the *GROWL*. L 27.________
28. The scarf was made of shiny *SILK*. H 28.________
29. The super highway has six *LANES*. H 29.________
30. He should know about the *HUT*. L 30.________

31. For dessert he had apple *PIE*. H 31.________
32. The beer drinkers raised their *MUGS*. H 32.________
33. I'm glad you heard about the *BEND*. L 33.________
34. You're talking about the *POND*. L 34.________
35. The rude remark made her *BLUSH*. H 35.________
36. Nancy had considered the *SLEEVES*. L 36.________
37. We heard the ticking of the *CLOCK*. H 37.________
38. He can't consider the *CRIB*. L 38.________
39. He killed the dragon with his *SWORD*. H 39.________
40. Tom discussed the *HAY*. L 40.________

41. Mary wore her hair in *BRAIDS*. H 41.________
42. She's glad Jane asked about the *DRAIN*. L 42.________
43. Bill hopes Paul heard about the *MIST*. L 43.________
44. We're lost so let's look at the *MAP*. H 44.________
45. No one was injured in the *CRASH*. H 45.________
46. We're speaking about the *TOLL*. L 46.________
47. My son has a dog for a *PET*. H 47.________
48. He was scared out of his *WITS*. H 48.________
49. We spoke about the *KNOB*. L 49.________
50. I've spoken about the *PILE*. L 50.________

Form #2 of the Revised SPIN Test

Name ________________________________(# ________) Marker ________ Date ________
S/B __+8 dB__ #C-HIGH ________ #C-LOW ________ ACCEPT? __Y / N__ Percent Hrg. ________

1.	Miss Black thought about the *LAP*.		L	1________
2.	The baby slept in his *CRIB*.	H		2________
3.	The watchdog gave a warning *GROWL*.	H		3________
4.	Miss Black would consider the *BONE*.		L	4________
5.	The natives built a wooden *HUT*.	H		5________
6.	Bob could have known about the *SPOON*.		L	6________
7.	Unlock the door and turn the *KNOB*.	H		7________
8.	He wants to talk about the *RISK*.		L	8________
9.	He heard they called about the *LANES*.		L	9________
10.	Wipe your greasy hands on the *RAG*.	H		10________
11.	She has known about the *DRUG*.		L	11________
12.	I want to speak about the *CRASH*.		L	12________
13.	The wedding banquet was a *FEAST*.	H		13________
14.	I should have considered the *MAP*.		L	14________
15.	Paul hit the water with a *SPLASH*.	H		15________
16.	The ducks swam around on the *POND*.	H		16________
17.	Ruth must have known about the *PIE*.		L	17________
18.	The man should discuss the *OX*.		L	18________
19.	Bob stood with his hands on his *HIPS*.	H		19________
20.	The cigarette smoke filled his *LUNGS*.	H		20________
21.	They heard I called about the *PET*.		L	21________
22.	The cushion was filled with *FOAM*.	H		22________
23.	Ruth poured the water down the *DRAIN*.	H		23________
24.	Bill cannot consider the *DEN*.		L	24________
25.	This nozzle sprays a fine *MIST*.	H		25________
26.	The sport shirt has short *SLEEVES*.	H		26________
27.	She hopes Jane called about the *CALF*.		L	27________
28.	Jane has a problem with the *COIN*.		L	28________
29.	She shortened the hem of her *SKIRT*.	H		29________
30.	Paul hopes she called about the *TANKS*.		L	30________
31.	The girl talked about the *GIN*.		L	31________
32.	The guests were welcomed by the *HOST*.	H		32________
33.	Mary should think about the *SWORD*.		L	33________
34.	Ruth could have discussed the *WITS*.		L	34________
35.	The ship's captain summoned his *CREW*.	H		35________
36.	You had a problem with a *BLUSH*.		L	36________
37.	The flood took a heavy *TOLL*.	H		37________
38.	The car drove off the steep *CLIFF*.	H		38________
39.	We have discussed the *STEAM*.		L	39________
40.	The policemen captured the *CROOK*.	H		40________
41.	The door was opened just a *CRACK*.	H		41________
42.	Tom is considering the *CLOCK*.		L	42________
43.	The sand was heaped in a *PILE*.	H		43________
44.	You should not speak about the *BRAIDS*.		L	44________
45.	Peter should speak about the *MUGS*.		L	45________
46.	Household goods are moved in a *VAN*.	H		46________
47.	He has a problem with the *OATH*.		L	47________
48.	Follow this road around the *BEND*.	H		48________
49.	Tom won't consider the *SILK*.		L	49________
50.	The farmer baled the *HAY*.	H		50________

Form #3 of the Revised SPIN Test

Name ____________________ (# ________) Marker ________ Date ________

S/B +8 dB #C-HIGH ________ #C-LOW ________ ACCEPT? Y / N Percent Hrg. ________

1.	Kill the bugs with this *SPRAY*.	H		1________
2.	Mr. White discussed the *CRUISE*.		L	2________
3.	How much can I buy for a *DIME*.	H		3________
4.	Miss White thinks about the *TEA*.		L	4________
5.	We shipped the furniture by *TRUCK*.	H		5________
6.	He is thinking about the *ROAR*.		L	6________
7.	She's spoken about the *BOMB*.		L	7________
8.	My TV has a twelve-inch *SCREEN*.	H		8________
9.	That accident gave me a *SCARE*.	H		9________
10.	You want to talk about the *DITCH*.		L	10________
11.	The king wore a golden *CROWN*.	H		11________
12.	The girl swept the floor with a *BROOM*.	H		12________
13.	We're discussing the *SHEETS*.		L	13________
14.	The nurse gave him first *AID*.	H		14________
15.	She faced them with a foolish *GRIN*.	H		15________
16.	Betty has considered the *BARK*.		L	16________
17.	Watermelons have lots of *SEEDS*.	H		17________
18.	Use this spray to kill the *BUGS*.	H		18________
19.	Tom will discuss the *SWAN*.		L	19________
20.	The teacher sat on a sharp *TACK*.	H		20________
21.	You'd been considering the *GEESE*.		L	21________
22.	The sailor swabbed the *DECK*.	H		22________
23.	They were interested in the *STRAP*.		L	23________
24.	He could discuss the *BREAD*.		L	24________
25.	He tossed the drowning man a *ROPE*.	H		25________
26.	Jane hopes Ruth asked about the *STRIPES*.		L	26________
27.	Paul spoke about the *PORK*.		L	27________
28.	The boy gave the football a *KICK*.	H		28________
29.	The storm broke the sailboat's *MAST*.	H		29________
30.	Mr. Smith thinks about the *CAP*.		L	30________
31.	We are speaking about the *PRIZE*.		L	31________
32.	Mr. Brown carved the roast *BEEF*.	H		32________
33.	The glass had a chip on the *RIM*.	H		33________
34.	Harry had thought about the *LOGS*.		L	34________
35.	Bob could consider the *POLE*.		L	35________
36.	Her cigarette had a long *ASH*.	H		36________
37.	Ruth has a problem with the *JOINTS*.		L	37________
38.	He is considering the *THROAT*.		L	38________
39.	The soup was served in a *BOWL*.	H		39________
40.	We can't consider the *WHEAT*.		L	40________
41.	The man spoke about the *CLUE*.		L	41________
42.	The lonely bird searched for its *MATE*.	H		42________
43.	Please wipe your feet on the *MAT*.	H		43________
44.	David has discussed the *DENT*.		L	44________
45.	The pond was full of croaking *FROGS*.	H		45________
46.	He hit me with a clenched *FIST*.	H		46________
47.	Bill heard Tom called about the *COACH*.		L	47________
48.	A bicycle has two *WHEELS*.	H		48________
49.	Jane has spoken about the *CHEST*.		L	49________
50.	Mr. White spoke about the *FIRM*.		L	50________

Form #4 of the Revised SPIN Test

Name ____________________________(# ______) Marker ______ Date ______
S/B __+8 dB__ #C-HIGH ______ #C-LOW ______ ACCEPT? __Y / N__ Percent Hrg. ______

1.	The doctor X-rayed his *CHEST*.	H		1.__________
2.	Mary had considered the *SPRAY*.		L	2.__________
3.	The woman talked about the *FROGS*.		L	3.__________
4.	The workers are digging a *DITCH*.	H		4.__________
5.	Miss Brown will speak about the *GRIN*.		L	5.__________
6.	Bill can't have considered the *WHEELS*.		L	6.__________
7.	The duck swam with the white *SWAN*.	H		7.__________
8.	Your knees and your elbows are *JOINTS*.	H		8.__________
9.	Mr. Smith spoke about the *AID*.		L	9.__________
10.	He hears she asked about the *DECK*.		L	10.__________
11.	Raise the flag up the *POLE*.	H		11.__________
12.	You want to think about the *DIME*.		L	12.__________
13.	You've considered the *SEEDS*.		L	13.__________
14.	The detectives searched for a *CLUE*.	H		14.__________
15.	Ruth's grandmother discussed the *BROOM*.		L	15.__________
16.	The steamship left on a *CRUISE*.	H		16.__________
17.	Miss Smith considered the *SCARE*.		L	17.__________
18.	Peter has considered the *MAT*.		L	18.__________
19.	Tree trunks are covered with *BARK*.	H		19.__________
20.	The meat from a pig is called *PORK*.	H		20.__________
21.	The old man considered the *KICK*.		L	21.__________
22.	Ruth poured herself a cup of *TEA*.	H		22.__________
23.	We saw a flock of wild *GEESE*.	H		23.__________
24.	Paul could not consider the *RIM*.		L	24.__________
25.	How did your car get that *DENT*?	H		25.__________
26.	She made the bed with clean *SHEETS*.	H		26.__________
27.	I've been considering the *CROWN*.		L	27.__________
28.	The team was trained by their *COACH*.	H		28.__________
29.	I've got a cold and a sore *THROAT*.	H		29.__________
30.	We've spoken about the *TRUCK*.		L	30.__________
31.	She wore a feather in her *CAP*.	H		31.__________
32.	The bread was made from whole *WHEAT*.	H		32.__________
33.	Mary could not discuss the *TACK*.		L	33.__________
34.	Spread some butter on your *BREAD*.	H		34.__________
35.	The cabin was made of *LOGS*.	H		35.__________
36.	Harry might consider the *BEEF*.		L	36.__________
37.	We're glad Bill heard about the *ASH*.		L	37.__________
38.	The lion gave an angry *ROAR*.	H		38.__________
39.	The sandal has a broken *STRAP*.	H		39.__________
40.	Nancy should consider the *FIST*.		L	40.__________
41.	He's employed by a large *FIRM*.	H		41.__________
42.	They did not discuss the *SCREEN*.		L	42.__________
43.	Her entry should win first *PRIZE*.	H		43.__________
44.	The old man thinks about the *MAST*.		L	44.__________
45.	Paul wants to speak about the *BUGS*.		L	45.__________
46.	The airplane dropped a *BOMB*.	H		46.__________
47.	You're glad she called about the *BOWL*.		L	47.__________
48.	A zebra has black and white *STRIPES*.	H		48.__________
49.	Miss Black could have discussed the *ROPE*.		L	49.__________
50.	I hope Paul asked about the *MATE*.		L	50.__________

Form #5 of the Revised SPIN Test

Name ________________________________(# ________) Marker ________ Date ________
S/B +8 dB #C-HIGH ________ #C-LOW ________ ACCEPT? Y / N Percent Hrg. ________

1.	Betty knew about the *NAP.*		L	1________
2.	The girl should consider the *FLAME.*		L	2________
3.	It's getting dark, so light the *LAMP.*	H		3________
4.	To store his wood he built a *SHED.*	H		4________
5.	They heard I asked about the *BET.*		L	5________
6.	The mouse was caught in the *TRAP.*	H		6________
7.	Mary knows about the *RUG.*		L	7________
8.	The airplane went into a *DIVE.*	H		8________
9.	The fireman heard her frightened *SCREAM.*	H		9________
10.	He was interested in the *HEDGE.*		L	10________
11.	He wiped the sink with a *SPONGE.*	H		11________
12.	Jane did not speak about the *SLICE.*		L	12________
13.	Mr. Brown can't discuss the *SLOT.*		L	13________
14.	The papers were held by a *CLIP.*	H		14________
15.	Paul can't discuss the *WAX.*		L	15________
16.	Miss Brown shouldn't discuss the *SAND.*		L	16________
17.	The chicks followed the mother *HEN.*	H		17________
18.	David might consider the *FUN.*		L	18________
19.	She wants to speak about the *ANT.*		L	19________
20.	The fur coat was made of *MINK.*	H		20________
21.	The boy took shelter in a *CAVE.*	H		21________
22.	He hasn't considered the *DART.*		L	22________
23.	Eve was made from Adam's *RIB.*	H		23________
24.	The boat sailed along the *COAST.*	H		24________
25.	We've been discussing the *CRATES.*		L	25________
26.	The judge is sitting on the *BENCH.*	H		26________
27.	We've been thinking about the *FAN.*		L	27________
28.	Jane didn't think about the *BROOK.*		L	28________
29.	Cut a piece of meat from the *ROAST.*	H		29________
30.	Betty can't consider the *GRIEF.*		L	30________
31.	The heavy rains caused a *FLOOD.*	H		31________
32.	The swimmer dove into the *POOL.*	H		32________
33.	Harry will consider the *TRAIL.*		L	33________
34.	Let's invite the whole *GANG.*	H		34________
35.	The house was robbed by a *THIEF.*	H		35________
36.	Tom is talking about the *FEE.*		L	36________
37.	Bob wore a watch on his *WRIST.*	H		37________
38.	Tom had spoken about the *PILL.*		L	38________
39.	Tom has been discussing the *BEADS.*		L	39________
40.	The secret agent was a *SPY.*	H		40________
41.	The rancher rounded up his *HERD.*	H		41________
42.	Tom could have thought about the *SPORT.*		L	42________
43.	Mary can't consider the *TIDE.*		L	43________
44.	Ann works in the bank as a *CLERK.*	H		44________
45.	A champanzee is an *APE.*	H		45________
46.	He hopes Tom asked about the *BAR.*		L	46________
47.	We could discuss the *DUST.*		L	47________
48.	The bandits escaped from *JAIL.*	H		48________
49.	Paul hopes we heard about the *LOOT.*		L	49________
50.	The landlord raised the *RENT.*	H		50________

Form #6 of the Revised SPIN Test

Name ________________________________ (# ________) Marker ________ Date ________
S/B +8 dB #C-HIGH ________ #C-LOW ________ ACCEPT? Y / N Percent Hrg. ________

1.	You were considering the *GANG*.		L	1.________
2.	The boy considered the *MINK*.		L	2.________
3.	Playing checkers can be *FUN*.	H		3.________
4.	The doctor charged a low *FEE*.	H		4.________
5.	He wants to know about the *RIB*.		L	5.________
6.	The gambler lost the *BET*.	H		6.________
7.	Get the bread and cut me a *SLICE*.	H		7.________
8.	She might have discussed the *APE*.		L	8.________
9.	The sleepy child took a *NAP*.	H		9.________
10.	Instead of a fence, plant a *HEDGE*.	H		10.________
11.	The old woman discussed the *THIEF*.		L	11.________
12.	Drop the coin through the *SLOT*.	H		12.________
13.	They fished in the babbling *BROOK*.	H		13.________
14.	You were interested in the *SCREAM*.		L	14.________
15.	We hear they, asked about the *SHED*.		L	15.________
16.	The widow's sob expressed her *GRIEF*.	H		16.________
17.	The candle flame melted the *WAX*.	H		17.________
18.	I haven't discussed the *SPONGE*.		L	18.________
19.	He was hit by a poisoned *DART*.	H		19.________
20.	Ruth had a necklace of glass *BEADS*.	H		20.________
21.	Ruth will consider the *HERD*.		L	21.________
22.	The singer was mobbed by her *FANS*.	H		22.________
23.	The old man discussed the *DIVE*.		L	23.________
24.	The class should consider the *FLOOD*.		L	24.________
25.	The fruit was shipped in wooden *CRATES*.	H		25.________
26.	I'm talking about the *BENCH*.		L	26.________
27.	Paul has discussed the *LAMP*.		L	27.________
28.	The candle burned with a bright *FLAME*.	H		28.________
29.	You knew about the *CLIP*.		L	29.________
30.	She might consider the *POOL*.		L	30.________
31.	We swam at the beach at high *TIDE*.	H		31.________
32.	Bob was considering the *CLERK*.		L	32.________
33.	We got drunk in the local *BAR*.	H		33.________
34.	A termite looks like an *ANT*.	H		34.________
35.	The man knew about the *SPY*.		L	35.________
36.	The sick child swallowed the *PILL*.	H		36.________
37.	The class is discussing the *WRIST*.		L	37.________
38.	The burglar escaped with the *LOOT*.	H		38.________
39.	They hope he heard about the *RENT*.		L	39.________
40.	Mr. White spoke about the *JAIL*.		L	40.________
41.	He rode off in a cloud of *DUST*.	H		41.________
42.	Miss Brown might consider the *COAST*.		L	42.________
43.	Bill didn't discuss the *HEN*.		L	43.________
44.	The bloodhound followed the *TRAIL*.	H		44.________
45.	The boy might consider the *TRAP*.		L	45.________
46.	On the beach we play in the *SAND*.	H		46.________
47.	He should consider the *ROAST*.		L	47.________
48.	Miss Brown spoke about the *CAVE*.		L	48.________
49.	She hated to vacuum the *RUG*.	H		49.________
50.	Football is a dangerous *SPORT*.	H		50.________

Form #7 of the Revised SPIN Test

Name ______________________________(# ________) Marker ________ Date ________
S/B __+8 dB__ #C-HIGH ________ #C-LOW ________ ACCEPT? __Y / N__ Percent Hrg. ________

1.	We're considering the *BROW*.		L	1.____________
2.	You cut the wood against the *GRAIN*.	H		2.____________
3.	I am thinking about the *KNIFE*.		L	3.____________
4.	They've considered the *SHEEP*.		L	4.____________
5.	The cop wore a bullet-proof *VEST*.	H		5.____________
6.	He's glad we heard about the *SKUNK*.		L	6.____________
7.	His pants were held up by a *BELT*.	H		7.____________
8.	Paul took a bath in the *TUB*.	H		8.____________
9.	The girl should not discuss the *GOWN*.		L	9.____________
10.	Maple syrup is made from *SAP*.	H		10.____________
11.	Mr. Smith knew about the *BAY*.		L	11.____________
12.	They played a game of cat and *MOUSE*.	H		12.____________
13.	The thread was wound on a *SPOOL*.	H		13.____________
14.	We did not discuss the *SHOCK*.		L	14.____________
15.	The crook entered a guilty *PLEA*.	H		15.____________
16.	Mr. Black has discussed the *CARDS*.		L	16.____________
17.	A bear has a thick coat of *FUR*.	H		17.____________
18.	Mr. Black considered the *FLEET*.		L	18.____________
19.	To open the jar, twist the *LID*.	H		19.____________
20.	We are considering the *CHEERS*.		L	20.____________
21.	Sue was interested in the *BRUISE*.		L	21.____________
22.	Tighten the belt by a *NOTCH*.	H		22.____________
23.	The cookies were kept in a *JAR*.	H		23.____________
24.	Miss Smith couldn't discuss the *ROW*.		L	24.____________
25.	I am discussing the *TASK*.		L	25.____________
26.	The marksman took careful *AIM*.	H		26.____________
27.	I ate a piece of chocolate *FUDGE*.	H		27.____________
28.	Paul should know about the *NET*.		L	28.____________
29.	Miss Smith might consider the *SHELL*.		L	29.____________
30.	John's front tooth had a *CHIP*.	H		30.____________
31.	At breakfast he drank some *JUICE*.	H		31.____________
32.	You cannot have discussed the *GREASE*.		L	32.____________
33.	I did not know about the *CHUNKS*.		L	33.____________
34.	Our cat is good at catching *MICE*.	H		34.____________
35.	I should have known about the *GUM*.		L	35.____________
36.	Mary hasn't discussed the *BLADE*.		L	36.____________
37.	The stale bread was covered with *MOLD*.	H		37.____________
38.	Ruth has discussed the *PEG*.		L	38.____________
39.	How long can you hold your *BREATH*?	H		39.____________
40.	His boss made him work like a *SLAVE*.	H		40.____________
41.	We have not thought about the *HINT*.		L	41.____________
42.	Air mail requires a special *STAMP*.	H		42.____________
43.	The bottle was sealed with a *CORK*.	H		43.____________
44.	The old man discussed the *YELL*.		L	44.____________
45.	They're glad we heard about the *TRACK*.		L	45.____________
46.	Cut the bacon into *STRIPS*.	H		46.____________
47.	Throw out all this useless *JUNK*.	H		47.____________
48.	The boy can't talk about the *THORNS*.		L	48.____________
49.	Bill won't consider the *BRAT*.		L	49.____________
50.	The shipwrecked sailors built a *RAFT*.	H		50.____________

Form #8 of the Revised SPIN Test

Name ______________________________ (# ________) Marker ________ Date ________

S/B +8 dB #C-HIGH ________ #C-LOW ________ ACCEPT? Y / N Percent Hrg. ________

1.	Bob heard Paul called about the *STRIPS*.		L	1.________
2.	My turtle went into its *SHELL*.	H		2.________
3.	Paul has a problem with the *BELT*.		L	3.________
4.	I cut my finger with a *KNIFE*.	H		4.________
5.	They knew about the *FUR*.		L	5.________
6.	We're glad Ann asked about the *FUDGE*.		L	6.________
7.	Greet the heroes with loud *CHEERS*.	H		7.________
8.	Jane was interested in the *STAMP*.		L	8.________
9.	That animal stinks like a *SKUNK*.		L	9.________
10.	A round hole won't take a square *PEG*.	H		10.________
11.	Miss White would consider the *MOLD*.		L	11.________
12.	They want to know about the *AIM*.		L	12.________
13.	The admiral commands the *FLEET*.	H		13.________
14.	The bride wore a white *GOWN*.	H		14.________
15.	The woman discussed the *GRAIN*.		L	15.________
16.	You hope they asked about the *VEST*.		L	16.________
17.	I can't guess so give me a *HINT*.	H		17.________
18.	Our seats were in the second *ROW*.	H		18.________
19.	We should have considered the *JUICE*.		L	19.________
20.	The boat sailed across the *BAY*.	H		20.________
21.	The woman considered the *NOTCH*.		L	21.________
22.	That job was an easy *TASK*.	H		22.________
23.	The woman knew about the *LID*.		L	23.________
24.	Jane wants to speak about the *CHIP*.		L	24.________
25.	The shepherd watched his flock of *SHEEP*.	H		25.________
26.	Bob should not consider the *MICE*.		L	26.________
27.	David wiped the sweat from his *BROW*.	H		27.________
28.	Ruth hopes she called about the *JUNK*.		L	28.________
29.	I can't consider the *PLEA*.		L	29.________
30.	The bad news came as a *SHOCK*.	H		30.________
31.	A spoiled child is a *BRAT*.	H		31.________
32.	Paul was interested in the *SAP*.		L	32.________
33.	The drowning man let out a *YELL*.	H		33.________
34.	A rose bush has prickly *THORNS*.	H		34.________
35.	He's glad you called about the *JAR*.		L	35.________
36.	The dealer shuffled the *CARDS*.	H		36.________
37.	Miss Smith knows about the *TUB*.		L	37.________
38.	The man would not discuss the *MOUSE*.		L	38.________
39.	The railroad train ran off the *TRACK*.	H		39.________
40.	My jaw aches when I chew *GUM*.	H		40.________
41.	Ann was interested in the *BREATH*.		L	41.________
42.	You're glad they heard about the *SLAVE*.		L	42.________
43.	He caught the fish in his *NET*.	H		43.________
44.	Bob was cut by the jackknife's *BLADE*.	H		44.________
45.	The man could consider the *SPOOL*.		L	45.________
46.	Tom fell down and got a bad *BRUISE*.	H		46.________
47.	Lubricate the car with *GREASE*.	H		47.________
48.	Peter knows about the *RAFT*.		L	48.________
49.	Cut the meat into small *CHUNKS*.	H		49.________
50.	She hears Bob asked about the *CORK*.		L	50.________

Chapter 3

VALIDITY ISSUES IN SPEECH RECOGNITION TESTING

BRIAN E. WALDEN

Walter Reed Army Medical Center
Washington, DC

In recent years, a group at Walter Reed Army Medical Center has become increasingly concerned with validating the test procedures that are used with hearing-impaired patients. Several current research projects deal with the validity problems associated with testing the hearing impaired. This paper will review the conceptual framework for validation research.

Fundamentally, a test procedure or other measuring instrument is valid if it measures what it is supposed to measure. The validity of some test instruments is rather easy to establish—for example, the sound level meter as a measure of sound pressure. The decibel readings obtained should be uniformly consistent with certain well-accepted rules governing the relationship between sound pressure and decibels sound pressure level. Unfortunately, the validity of many test instruments, particularly in the behavioral sciences, is not so easily verified. Many applications of the most popular measures of speech recognition are illustrative of this fact.

In order to establish the validity of a test instrument, some kind of empirical investigation is necessary. The nature of the evidence required to establish a test's validity depends on the type of validity sought by the developer of the test instrument. At least four types are relevant to a discussion of validity issues in speech recognition assessment. These are *predictive validity, content validity, construct validity,* and *face validity.*

Predictive validity must be established when a test instrument is used to estimate some criterion behavior. This criterion is a reflection of the basic purpose or purposes for testing. There are, for example, at least five purposes for the speech recognition testing that is performed within the military: *administrative* (that is, job placement and compensation); *differential diagnosis; estimating everyday communication ability and rehabilitative needs; hearing aid evaluations;* and *simple descriptive purposes in medical evaluations.* Each of these applications of speech recognition testing requires a different criterion behavior. The validity of a test is established for each application by correlating scores on the predictor test with performance on the appropriate criterion variable. The size of the correlation is a direct indication of the amount of predictive validity inherent in the test.

With the exception of using speech recognition tests for differential diagnosis (i.e., to predict site of lesion), the predictive validity of common speech discrimination tests is relatively unknown. Jerger and Jerger (1971), Dirks, Kamm, Bower, and Betsworth (1977), and Bess, Josey, and Humes (1979) have demonstrated that performance-intensity functions, based on monosyllabic word recognition tests, provide a relatively valid method of predicting retrocochlear lesions. Similar empirical investigations to establish the predictive validity of common speech recognition tests for other applications have been relatively rare and generally less definitive. The unknown extent to which scores on monosyllabic word recognition tests predict communication ability in everyday listening situations is illustrative. Yet, results from such tests are regularly interpreted by rehabilitative audiologists from this perspective. Similarly, the validity of these tests to predict performance on the job has relatively little empirical foundation.

The lack of appropriate predictive validation studies in speech recognition testing owes principally to difficulty in quantifying the criterion behaviors. Only in the application of speech identification tests to predicting site of lesion is the criterion variable well defined—that is, confirmed organic pathology. In such a case, the predictive validity of a variety of tests can be compared and the test having the greatest validity selected for continued use. For the other applications of speech recognition testing, however, the criterion behaviors have not been empirically defined with precision. There is no single, satisfactory index of everyday communication ability, for example. For most applications of speech recognition testing, the criterion behavior is multidimentional and dynamic. These characteristics make quantification extremely complex and, realistically, in some cases perhaps nearly impossible.

At least for the time being, it may be necessary to develop alternative validation criteria to the most intuitive ones in order to conduct the necessary predictive validation studies. This can be illustrated by considering the application of speech recognition testing to predict performance with a hearing aid in everyday listening. One of the assumptions underlying the traditional comparative hearing-aid evaluation is that relative performance with a set of aids predicts relative performance in everyday communication. One might wish to compare the predictive validity of two test instruments such as the NU-6 test (Tillman & Carhart, 1966) and the SSI (Speaks, 1967) for this purpose. Lacking an empirical quantitative index of the intuitive criterion behavior (i.e., everyday success with the aids), a simpler criterion might be used.

In recent research at Walter Reed, for example, investigators have used (a) the frequency of hearing aid use, (b) the frequency with which major changes in the fitting are required following the clinical evaluation, (c) the specific listening situations in which amplification is subjectively viewed to provide the most benefit, and (d) the patient acceptability ratings as alternative validation criteria in predictive validation studies of hearing aid evaluation procedures. Although the use of such alternative criterion variables may not be totally satisfactory, given that each is an indirect measure of everyday success with amplification, it is a clear step ahead of ignoring predictive validity all together.

Another approach to validating a test method is to establish its content validity. In this case the test is used to measure directly the behavior in a content domain rather than to predict some criterion behavior. That is, performance on the test itself is the behavior of interest. The extent to which the test provides a representative sample of a particular content domain determines the validity of the test and not the correlation of the test scores with another variable. Content validity, therefore, is provided by the procedures that are followed during the original construction of the test. The items on the test are not selected at random from the content domain, but rather, according to a thoroughly developed plan for test construction designed to provide a representative set of items from the content domain. The content stressed on the test is that which the test developer considers most essential to the behavior being measured.

Historically, it would appear that content validity has been given considerable attention in the development of several well-known tests of speech recognition. The development of the PB-50 word lists (Egan, 1948) for evaluating communication systems is illustrative. The domain of interest was monosyllabic words of the English language. The sampling plan for selecting items and constructing test lists from this domain called for the following: The lists must be of equal average difficulty, each list must have a composition representative of English speech (i.e., they must be "phonetically balanced"), and the words must be in common usage. Additional illustrations of attention to content validity in test construction are provided by the W-22 (Hirsh et al., 1952), CNC (Peterson & Lehiste, 1962), and NU-6 tests. These differ from the PB-50 word lists primarily in the emphasis that was placed on various elements of the content domain. In other cases, such as the PAL-8 (Egan, 1948) sentence test, a totally different domain of interest was sampled. In any case, the test may be considered to possess content validity to the extent that the items sampled represent the domain of interest and the domain sampled is appropriate to the purpose for which the test is being used. Currently, a group at Walter Reed is developing a self-assessment communication inventory for use with the hearing impaired. Content validity has played a major role in test construction and item selection.

A third type of validity relevant to a discussion of speech recognition testing is construct validity, which is an issue when the variable being measured is abstract rather than concrete. Such variables are constructs that do not exist as isolated, observable dimensions of behavior. In the case of speech recognition assessment, the construct of interest is represented by such terms as *speech discrimination, speech identification,* and *speech recognition.* Like classic constructs such as *anxiety* or *intelligence,* speech discrimination is an abstract variable rather than a specific, observable variable.

It is unlikely that many popular tests of speech recognition were developed as measures of this construct. Rather, the designers of these instruments intended them as measures of monosyllabic word recognition, consonant recognition, sentence recognition, and so on. When applied clinically, however, the original purpose of the test as a measure of an isolated, specific variable frequently has been lost and, from the clinician's perspective, the test is considered a measure of the construct *speech discrimination.* Audiologists regularly speak of a patient's speech discrimination ability, for example, without reference to the specific variable that was actually measured.

Because no single observable can perfectly represent an abstract construct, the best overall measure is obtained by combining test results from several behavioral domains, each of which is thought to reflect the underlying construct. The construct validity of any one of these individual indices is indicated by the extent to which the results of that test represent the overall results obtained when all of the observables in the domain are measured.

After specifying the domain of variables, the principal method by which the construct can be validated is determining to what extent all or some of those observables correlate with each other and are affected alike by experimental treatments or individual differences. Evidence for the validity of those popular tests, frequently interpreted to measure the construct *speech discrimination,* suggest that generally low correlations exist among scores from various measures of speech recognition such as nonsense syllables, monosyllables, sentences, and continuous discourse (Giolas & Epstein, 1963; Owens & Schubert, 1977; Speaks, Jerger, & Trammell, 1970; Williams & Hecker, 1968). It would appear, therefore, that more than one construct is measured by these variables and better specification of the domain of observables relating to the construct is required. In the meantime, the constructs of *speech discrimination* or *speech recognition* cannot be used with precision.

A fourth type of validity which is frequently referred to is face validity. This is a rather weak form of validity which concerns the extent to which a test instrument appears to measure what it is supposed to measure. Thus, for example, a sentence identification task presented in a multitalker babble has more face validity as an index of speech recognition ability at a cocktail party than, for example, a nonsense syllable test presented in quiet. Although it would appear that the sentence identification task would be a more valid measure of the criterion behavior than the nonsense syllable task, there is no guarantee that this is the case. A test instrument which has relatively good content validity will also exhibit face validity, since face validity is a part of content validity. Recall that content validity is ensured by the procedures for test construction. The judgment of face validity is made at the completion of test construction to determine if the plan of content appears to have been successful. The determination of face validity, therefore, is a global judgment of the completed test regarding the extent to which it appears to measure what it is supposed to measure.

When a test is intended to predict a criterion behavior, face validity is essentially irrelevant, because the validity of the test depends entirely on the correlation between the test scores and the criterion variable. Therefore, face validity must be considered minimally important to the development and evaluation of speech recognition tests other than as an aid in formulating hypotheses about possible test measures that may correlate positively with the criterion behavior. Most certainly, face validity is neither a necessary nor sufficient standard for establishing the validity of speech recognition tests as they are typically used.

When the state-of-the-art of speech recognition testing is considered in relation to the four types of validity that have been discussed, it would appear that the emphasis of test developers has been on content validity and face validity. In contrast, users of these tests often tend to think of them as measures of a construct, thereby making the construct validity of these instruments a relevant issue. Despite this, almost all applications of speech recognition testing are for the purpose of prediction. Given the basic definition of validity—that a test instrument is valid if it measures what it is supposed to measure—definitive validation studies of these tests must establish their predictive validity. Such studies are extremely difficult to design and conduct. It may be necessary to compromise on the selection of the criterion behavior in conducting these studies. Until empirical evidence of the predictive validity of various speech recognition tests are available, however, their use for predictive purposes will remain susceptible to criticism.

In order to establish the validity of existing tests of speech recognition and to develop new assessment methods, three questions must be answered. First, what are the most reasonable validity criteria for the various applications of speech recognition testing and how should each criterion behavior be operationally defined? This question recognizes that predictive validity must be the focus of attention, given current applications of speech recognition testing. The difficulty, obviously, is in how to define and quantify the criterion behavior. This issue would appear to be the most fundamental of all validity problems in speech recognition assessment.

Second, what is the most appropriate testing purpose for each of the various test materials currently in use (i.e., nonsense syllables, words, sentences, etc.)? It would appear that these measures are not observables from the same behavioral domain and that each is sensitive to a different aspect of the listener's ability to understand speech. These domains need to be better specified. Obviously, if there is uncertainty about exactly what it is that our test materials reflect, intelligent selection of materials for specific testing purposes is difficult.

Finally, what is a "clinically significant difference" for each of the various applications of speech recognition testing? Can it be reasonably defined in terms of test scores? Is it likely to vary in magnitude as a function of performance level or listening conditions? Fundamentally, the purpose of speech recognition testing is to detect a clinically significant difference in the domain of interest. Thus, if the purpose of testing is to determine compensation, the basic issue to be resolved is the magnitude of the performance difference on the test instrument that reflects a quantal difference in degree of handicap in everyday communication. Similarly, if the purpose of testing is hearing aid selection, it is necessary to know how large an interaid difference score is required before a meaningful performance difference can be expected between two instruments in everyday listening.

While the definition of a clinically significant difference for each of these behavioral domains is fundamentally a validity issue, it is also essential to a clinically meaningful definition of reliability. Reliability, like validity, is not an absolute attribute a test may possess. Rather, both reliability and validity must be assessed in relation to some external criterion. In the case of reliability, this is usually some statistical criterion. Such a criterion, however, indicates the probability that two test scores on a test instrument represent a numerically significant difference. Although this is a prerequisite to a clinically significant difference, it does not guarantee it. In speech recognition testing, therefore, an alternative criterion against which to assess the reliability of a test instrument is that the test-retest variability must be less than a clinically significant difference in the behavioral domain of interest. If not, the reliability of the test instrument must be considered inadequate for that clinical application. Such a practical definition of reliability, however, must wait for the definition of a clinically significant difference in each behavioral domain.

ACKNOWLEDGMENT

The opinions and assertions contained herein are the private views of the author and not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

REFERENCES

Bess, F. H., Josey, A. F., & Humes, L. E. (1979). Performance intensity functions in cochlear and eighth nerve disorders. *American Journal of Otolaryngology, 1,* 27–31.

Dirks, D., Kamm, C., Bower, D., & Betsworth, A. (1977). Use of performance-intensity functions for diagnosis. *Journal of Speech and Hearing Disorders, 42,* 408–415.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope, 58,* 955–981.

Giolas, T. G., & Epstein, A. (1963). Comparative intelligibility of word lists and continuous discourse. *Journal of Speech and Hearing Research, 6,* 349–358.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17,* 321–337.

Jerger, J., & Jerger, S. (1971). Diagnostic significance of PB word functions. *Archives of Otolaryngology, 93,* 573–580.

Owens, E., & Schubert, E. D. (1977). Development of the California Consonant Test. *Journal of Speech and Hearing Research, 20,* 463–474.

Peterson, G. E., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders, 27,* 62–70.

Speaks, C. (1967). Intelligibility of filtered synthetic sentences. *Journal of Speech and Hearing Research, 10,* 289–298.

Speaks, C., Jerger, J., & Trammell, J. (1970). Comparison of sentence identification and conventional speech discrimination scores. *Journal of Speech and Hearing Research, 13,* 755–767.

Tillman, T. W., & Carhart, R. (1966). *An expanded test for speech discrimination utilizing CNC monosyllabic words* (Tech. Rep. SAM-TR-66-55). Brooks AFB, TX: USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC).

Williams, C. E., & Hecker, M. H. L. (1968). Relation between intelligibility scores for four test methods and three types of speech distortion. *Journal of the Acoustical Society of America, 44,* 1002–1006.

Chapter 4

PSYCHOMETRIC ISSUES IN SPEECH RECOGNITION TESTING

MARILYN E. DEMOREST

University of Maryland Baltimore County

The assessment of individual differences in any aspect of human behavior involves two components: observation and inference. Assessment can be an informal process that begins with unstructured interaction between an observer and a subject and leads to a subjective clinical judgment about how this individual differs from others. Or, in contrast, it can refer to a highly structured, standardized, and systematic form of observation that involves objective quantification or measurement. Inferences are then drawn by interpreting the meaning or significance of the individual's scores.

When assessment involves formal measurement procedures, one is likely to be concerned about the psychometric properties of the obtained scores. Do the scores form an ordinal scale that can only rank individuals, or do they form an interval scale with a meaningful unit of measurement? To what extent can one generalize from the observed score and draw inferences about what is likely to be observed on other occasions or with other observers or with different but equivalent procedures? Can one make predictions about the individual's scores on other variables? Questions such as these are addressed by studying the statistical properties of the observed scores.

In assessing speech recognition in the hearing impaired, psychometric issues are relevant to the extent that scores on speech recognition tasks are expected to differ among individuals. When scores are used for diagnostic purposes, for selecting hearing aids, for estimating communication handicap, or for almost any other clinical purpose, it is clear that patients are expected to differ from one another, and therefore proper interpretation of scores will be facilitated if the psychometric properties of the scores are known.

Although few audiologists or speech-language pathologists would probably label their activities in assessing speech recognition as *psychological testing*, according to the *Standards for Educational & Psychological Tests* (see the Appendix) published by the American Psychological Association (1974, p. 2):

> A test is a special case of an assessment procedure. It may be thought of as a set of tasks or questions intended to elicit particular types of behavior when presented under standardized conditions and to yield scores that will have desirable psychometric properties.

Because most speech recognition testing is consistent with this definition, I think it may be instructive to consider briefly the kinds of standards which measurement specialists in education and psychology have set forth as important principles in the construction, administration, interpretation, and evaluation of tests. Not all of the standards are relevant in this context, but they should be considered and judged for their relevance, and should identify areas in which our future efforts might be directed.

The 1974 Standards are a revision of an earlier document published in 1966 by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. They are organized by content and by importance. In terms of content they range over several broad areas including the following:

A. Dissemination of Information
B. Aids to Interpretation
C. Directions for Administration and Scoring
D. Norms and Scales
E. Validity
F. Reliability and Measurement Error
G. Qualifications and Concerns of Users
H. Choice or Development of Test or Method
I. Administration and Scoring
J. Interpretation of Scores

In terms of importance, individual standards are designated as *Essential*, *Very Desirable*, or *Desirable*. The standards listed as Essential are intended to "represent the consensus of present-day thinking concerning what is normally required for competent use of a test" (APA, 1974, p. 6). The term *Very Desirable* is used "to draw attention to types of information or practices that contribute greatly to the user's understanding of the test and to competence in its use" (p. 7). Desirable information and practices are those that are "helpful but not Essential or Very Desirable" (p. 7). It should be noted that the sponsoring organizations put forth the standards for consideration by professionals in a spirit of voluntary self-evaluation. They are not written as law and individual competence is not to be judged in terms of the "literal satisfaction of every relevant provision" (p. 8). With these points in mind, an overview of the standards and comments on certain issues follows.

The first section of the Standards specifies that when a test is put forth for use by the testing community, it should be accompanied by a manual or similar document that describes the rationale underlying the test, its development, and any evidence in support of claims that are made for its use. Such manuals do not exist for tests routinely used in the clinic for adults. Although there is plentiful research on these tests, relevant information has not been compiled into easily accessible manuals or handbooks.

Under Aids to Interpretation it is stressed that the test user should be given as much information as possible that will aid in correct interpretation of test scores. It is particularly important to explicitly state the purposes for which the test is recommended. A speech test recommended for its diagnostic value, for example, might not be a good measure of communication handicap. Other aids include mention of any special user qualifications required to score and administer the test, the rationale underlying the test, and evidence in support of any claims that are made for reliability and validity.

Standards that concern Directions for Administration and Scoring are straightforward: They specify that sufficient information must be provided for the user to ensure standardization of procedure. Because a multitude of extraneous factors produces unwanted variability in observed scores, it is essential that procedures be standardized to the greatest extent possible.

Standards dealing with Norms and Scales present some interesting issues when testing is being conducted on a clinical population. When a test score can be interpreted directly in terms of item content, the score of an individual patient can be interpreted without reference to the scores of other individuals. If, for example, the test items represent a sample from a well-defined population of possible items (CV nonsense syllables, English monosyllables of a certain type, etc.), then performance on the sample items can be used to infer what performance would be on the population of items.

Frequently, however, there is a desire to compare an individual's score with the scores of others. This is usually done by expressing the score relative to the distribution of scores obtained by some reference group. When one is interested in detecting impairment, it seems appropriate to compare a patient's score with the distribution of scores obtained by normal-hearing listeners. If there are no true individual differences among normal-hearing listeners on the test, then any observed variability in their scores simply reflects variability due to measurement error, and when the patient's score differs sufficiently from the mean score for normal-hearing listeners, one can infer the presence of some impairment. The performance of normals serves as a standard against which to calibrate impairment. As the true variability among normal-hearing listeners grows, relative to the variability attributable to measurement error, the difficulty in detecting impairment relative to the normal group grows also. With two distinct subpopulations, however, it is possible that the test does not measure the same thing within each population. That is, there may be true individual differences among the normal-hearing listeners but along a dimension systematically different in some way from that which differentiates the hearing-impaired listeners. Expressing the patient's score in terms of the performance of a different population on a subtly different variable would then have less to recommend it.

An alternative strategy, and one which provides a different type of information, is to compare the individual's score to a reference group of which she/he is a member. The reference group might be similar in age, gender, degree or configuration of pure-tone loss, and so on. Expectancy tables can be constructed showing the distribution of test scores as a function of these other variables. The test score then provides the most information about the individual when his or her score deviates from this norm.

Another aspect of norming which deserves mention is the desirability of constructing local norms when feasible. If a particular clinic population is rather homogeneous or if it differs systematically in some way from the general population of hearing-impaired persons, then local norms can provide a stereotype of the typical patient. Again, the test score will provide the most information when it differs from this norm, that is, when the patient is not like the typical patient.

For example, at the Walter Reed Army Medical Center there is a relatively homogeneous population of hearing-impaired individuals. Most have high-frequency, noise-induced hearing loss. In such situations, there is a certain pattern that one expects to see, and only when a patient deviates from that pattern is further testing required. There is a potential for special subgroup norms and for special local norms to provide information about a stereotypical pattern that can be expected. Then, only when there is some evidence that this pattern is not present does a clinician need to pursue the testing to explore the case in more detail. The merits of this suggestion could be debated because it is very expensive to gather local norms, to maintain them, to update them, and so on.

Since Validity issues are already considered in some detail in another chapter of this report, the standards that deal with validity will not be enumerated except to point out that it is incumbent upon the test developer, or any other individual who promotes the use of a test for a particular purpose, to provide documentation of its usefulness for that purpose. Validity studies can be as diverse as the possible reasons for using the test and the possible inferences that will be drawn about its scores.

Regarding Reliability and Measurement Error, the Standards specify that evidence of reliability is essential so that the test user can determine whether the test is sufficiently dependable for its intended purpose. This evidence should include not only reliability coefficients but also estimates of the standard error of measurement. For a number of reasons, that latter statistic is preferable. Although the term *reliability* can in its generic sense refer simply to dependability or precision of measurement, reliability coefficients are statistics that quantify reliability for a given measurement procedure and a specific population or subpopulation. A reliability coefficient calculated within a sample of normal-hearing listeners may reveal little or no reliability if there is little or no true variability within that subpopulation. In contrast, the same test administered and scored in the same way may show extremely high reliability within a heterogeneous sample of hearing-impaired listeners. This is attributable to the greater true variability within that subpopulation. Reliability coefficients are useful for indicating how well individuals can be ranked within the population on the basis of their test scores.

For most clinical applications, reliability coefficients have limited usefulness. What is more important is a measure of how much scores are likely to fluctuate as a result of measurement error alone. This information is needed if one is to evaluate the significance of a difference in scores obtained at two points in time or under two different listening conditions, such as with and without a hearing aid. The standard error of measurement is the most useful statistic for this purpose. It must be remembered, however, that the standard error of measurement for a single score must be multiplied by a factor of square root of 2 if the difference between two scores is being assessed.

The section on Qualifications and Concerns of Users indicates that it is the responsibility of the test user to be informed about general measurement principles and also about the particular test being used. This is necessary to protect the patient from inappropriate interpretation and use of the test scores. Additional responsibilities of the test user involve the selection of the appropriate test method, adherence to the standard methods of administration and scoring, and competent score interpretation. At present, those who administer speech recognition tests are probably far more sophisticated in the principles of psychophysical measurement than in the principals and theories of psychological measurement. The availability of good test manuals and handbooks could do much to eliminate this discrepancy.

REFERENCE

AMERICAN PSYCHOLOGICAL ASSOCIATION. (1974). *Standards for educational & psychological tests.* Washington, DC: APA.

APPENDIX

Standards for Educational & Psychological Tests[1]

American Psychological Association, 1974

A. *Dissemination of Information*

1. When a test is published or otherwise made available for operational use, it should be accompanied by a manual (or other published or readily available information) that makes every reasonable effort to follow the recommendations of these standards and, in particular, to provide the information required to substantiate any claims that have been made for its use. (E)[2]

2. A test manual should describe fully the development of the test: the rationale, specifications followed in writing items or selecting observations, and procedures and results of item analysis or other research. (E)

3. The test and its manual should be revised at appropriate intervals. The time for revision has arrived whenever changing conditions of use or new research data make any statements in the manual incorrect or misleading. (VD)

B. *Aids to Interpretation*

1. The test, the manual, the record forms, and other accompanying material should help users make correct interpretations of the test results and should warn against common misuses. (E)

2. The test manual should state explicitly the purposes and applications for which the test is recommended. (E)

3. The test manual should describe clearly the psychological, educational, or other reasoning underlying the test and nature of the characteristic it is intended to measure. (E)

4. The test manual should identify any special qualifications required to administer the test and to interpret it properly. (E)

5. Evidence of validity and reliability, along with other relevant research data, should be presented in support of any claims being made. (E)

6. Test developers or others offering computer services for test interpretation should provide a manual reporting the rationale and evidence in support of computer-based interpretations of scores. (E)

C. *Directions for Administration and Scoring*

1. The directions for administration should be presented in the test manual with sufficient clarity and emphasis so that the test user can duplicate, and will be encouraged to duplicate, the administrative conditions under which the norms and the data on reliability and validity were obtained. (E)

2. Instructions should prepare the examinee for the examination: Sample material, practice use of answer sheets or punch cards, sample questions, etc., should be provided. (D)

3. The procedures for scoring the test should be presented in the test manual with a maximum of detail and clarity to reduce the likelihood of scoring error. (E)

D. *Norms and Scales*

1. Norms should be published in the test manual at the time of release of the test for operational use. (E)

2. Norms presented in the test manual should refer to defined and clearly described populations. These populations should be the groups with whom users of the test will ordinarily wish to compare the persons tested. (E)

3. In reporting norms, test manuals should use percentiles for one or more appropriate reference groups or standard scores for which the basis is clearly set forth; any exceptional type of score or unit should be explained and justified. Measures of central tendency and variability should always be reported. (E)

4. Local norms are more important for many uses of tests than are published norms. A test manual should suggest using local norms in such situations. (VD)

5. Derived scales used for reporting scores should be carefully described in the test manual to increase the likelihood of accurate interpretation of scores by both the test interpreter and the examinee. (E)

6. If scales are revised, new forms added, or other changes made, the revised test manual should provide tables of equivalence between the new and the old forms. This provision is particularly important in cases where data are recorded on cumulative records. (D)

7. Where it is expected that a test will be used to assess groups rather than individuals (i.e., for schools or programs), normative data based on group summary statistics should be provided. (E)

E. *Validity*

1. A manual or research report should present the evidence of validity for each type of inference for which use of the test is recommended. If validity for some suggested interpretation has not been investigated, that fact should be made clear. (E)

2. A test user is responsible for marshalling the evidence in support of his claims of validity and reliability. The use of test scores in decision rules should be supported by evidence. (E)

3. All measures of criteria should be described completely and accurately. The manual or research report should comment on the adequacy of a criterion. Whenever feasible, it should draw attention to

[1] Copyright 1974 by the American Psychological Association. Adapted by permission.

[2] The designation (E) indicates that consideration of the standard is deemed Essential. The statements listed as essential are intended to represent the consensus of present-day thinking concerning what is normally required for competent use of a test.

The designation (VD), or Very Desirable, is used to draw attention to types of information or practices that contribute greatly to the user's understanding of the test and to competence in its use.

The designation (D), or Desirable, includes information and practices that are helpful but not Essential or Very Desirable.

significant aspects of performance that the criterion measure does not reflect and to irrelevant factors likely to affect it. (E)

4. A criterion measure should itself be studied for evidence of validity and that evidence should be presented in the manual or report. (VD)

5. The manual or research report should provide information on the appropriateness of or limits to the generalizability of validity information. (VD)

6. The sample employed in a validity study and the conditions under which testing is done should be consistent with recommended test use and should be described sufficiently for the reader to judge its pertinence to his situation. (E)

7. The collection of data for a validity study should follow procedures consistent with the purposes of the study. (E)

8. Any statistical analysis of criterion-related validity should be reported in the manual in a form that enables the reader to determine how much confidence is to be placed in judgments or predictions regarding the individual. (E)

9. A test user should investigate the possibility of bias in tests or in test *items*. Wherever possible, there should be an investigation of possible differences in criterion-related validity for ethnic, sex, or other subsamples that can be identified when the test is given. The manual or research report should give the results for each subsample separately or report that no differences were found. (E)

10. When a scoring key, the selection of items, or the weighting of tests is based on one sample, the manual should report validity coefficients based on data obtained from one or more independent crossvalidation samples. Validity statements should not be based on the original sample. (E)

11. To the extent feasible, a test user who intends to continue employing a test over a long period of time should develop procedures for gathering data for continued research. (D)

12. If test performance is to be interpreted as a representative sample of performance in a universe of situations, the test manual should give a clear definition of the universe represented and describe the procedures followed in the sampling from it. (E)

13. If the author proposes to interpret scores on a test as measuring a theoretical variable (ability, trait, or attitude), his proposed interpretation should be fully stated. His theoretical construct should be distinguished from interpretations arising on the basis of other theories. (E)

F. *Reliability and Measurement Error*

1. The test manual or research report should present evidence of reliability, including estimates of the standard error of measurement, that permits the reader to judge whether scores are sufficiently dependable for the intended uses of the test. If any of the necessary evidence has not been collected, the absence of such information should be noted. (E)

2. The procedures and samples used to determine reliability coefficients or standard errors of measurement should be described sufficiently to permit a user to judge the applicability of the data reported to the individuals or groups with which he is concerned. (E)

3. Reports of reliability studies should ordinarily be expressed in the test manual in terms of variances of error components, standard errors of measurement, or product-moment reliability coefficients. Unfamiliar expressions of data should be clearly described, with references to their development. (E)

4. If two or more forms of a test are published for use with the same examinees, information on means, variances, and characteristics of items in the forms should be reported in the test manual along with the coefficients of correlation among their scores. If necessary evidence is not provided, the test manual should warn the reader against assuming equivalence of scores. (E)

5. Evidence of internal consistency should be reported for any unspeeded test. (VD)

6. The test manual should indicate to what extent test scores are stable, that is, how nearly constant the scores are likely to be if a parallel form of a test is administered after time has elapsed. The manual should also describe the effect of any such variation on the usefulness of the test. The time interval to be considered depends on the nature of the test and on what interpretation of the test scores is recommended. (E)

G. *Qualifications and Concerns of Users*

1. A test user should have a general knowledge of measurement principles and of the limitations of test interpretations. (E)

2. A test user should know and understand the literature relevant to the test he uses and the testing problems with which he deals. (VD)

3. One who has the responsibility for decisions about individuals or policies that are based on test results should have an understanding of psychological or educational measurement and of validation and other test research. (E)

4. Test users should seek to avoid bias in test selection, administration, and interpretation; they should try to avoid even the appearance of discriminatory practice. (E)

5. Institutional test users should establish procedures for periodic internal review of test use. (E)

H. *Choice or Development of Test or Method*

1. The choice or development of tests, test batteries, or other assessment procedures should be based on clearly formulated goals and hypotheses. (E)

2. A test user should consider more than one variable for assessment and the assessment of any given variable by more than one method. (E)

3. In choosing an existing test, a test user should relate its history of research and development to his intended use of the instrument. (E)

4. In general a test user should try to choose or to develop an assessment technique in which "tester-effect" is minimized, or in which reliability of assessment across testers can be assured. (E)

5. Test scores used for selection or other administrative decisions about an individual may not be useful for individual or program evaluation and vice versa. (D)

I. *Administration and Scoring*

1. A test user is expected to follow carefully the standardized procedures described in the manual for administering a test. (E)

2. The test administrator is responsible for establishing conditions, consistent with the principle of standardization, that enable each examinee to do his best. (E)

3. A test user is responsible for accuracy in scoring, checking, coding or recording test results. (E)

4. If specific cutting scores are to be used as a basis for decisions, a test user should have a rationale, justification, or explanation of the cutting scores adopted. (E)

5. The test user shares with the test developer or distributor a responsibility for maintaining test security. (E)

J. *Interpretation of Scores*

1. A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances. (E)

2. Test scores should ordinarily be reported only to people who are qualified to interpret them. If scores are reported, they should be accompanied by explanations sufficient for the recipient to interpret them correctly.

3. The test user should recognize that estimates of reliability do not indicate criterion-related validity. (E)

4. A test user should examine carefully the rationale and validity of computer-based interpretations of test scores. (E)

5. In norm-referenced interpretations, a test user should interpret an obtained score with reference to sets of norms appropriate for the individual tested and for the intended use. (E)

6. Any content-referenced interpretation should clearly indicate the domain to which one can generalize. (E)

7. The test user should consider alternative interpretations of a given score. (E)

8. The test user should be able to interpret test performance relative to other measures. (VD)

9. A test user should develop procedures for systematically eliminating from data files test-score information that has, because of the lapse of time, become obsolete. (E)

Chapter 5

SPEECH RECOGNITION TESTING AND UNDERSTANDING THE EFFECTS OF DISEASE ON FUNCTION OF THE EAR

AARON R. THORNTON

Massachusetts Eye and Ear Infirmary, Boston

In this discussion of clinical tests or services, all references will be to the situation in which the tests or services are performed with the intention of directly benefiting the patient or client in some way. With clinical services, one generally attempts to obtain only that information which will be useful in serving the client. The actual information acquired often rests with the judgment of the clinician, and the merits of having done more or less can always be argued. The point should be stressed that for clinical testing, cost/benefit ratio must always be evaluated. The more complicated and expensive tests must be justified by proportionately greater demonstrable benefit to the recipient. If a patient is known to have a specific lesion, then a test to demonstrate that lesion is often superfluous. Similarly, if a test localizes a lesion in a case where there is no possibility of intervention, or if the test provides no direction for intervention, then the need for testing may be questioned.

Before discussing the clinical uses of speech recognition testing, the clinical needs of the patients should be defined first and then assessed to determine which of them might best be met by speech recognition testing. Primary effort should be directed to the improvement of the health and welfare of patients. It is possible that in some cases this might be accomplished by eliminating the test rather than changing it. Clinicians who cannot embrace that concept have become disciples of a mode of testing and cannot objectively evaluate their work. The clinical evaluation needs of patients can be grouped under three headings: (a) diagnosis of disease, (b) localization and definition of lesions, and (c) assessment of functional impairment. The following sections present a somewhat skeptical view of speech recognition testing.

For the diagnosis of disease, speech recognition testing plays only a minor role. In the case of ear disease, the majority of treatable diseases involve middle ear lesions, for which speech testing provides little practical benefit. The etiology of sensorineural hearing loss is seldom determined conclusively; effective treatment is known for only a few diseases; and speech testing is virtually unnecessary in reaching a diagnosis. It is neither the least expensive, the most sensitive, nor the most accurate diagnostic measure for any type of ear disease. Speech recognition testing is equally ineffective for localizing and defining lesions of the ear. For lesions of the central nervous system, relatively few cases can be found in which the medical management or rehabilitative follow-up would have differed significantly if speech recognition testing had not been done for the purpose of localizing or defining a lesion.

On a more positive note, speech recognition testing clearly plays an important role in the assessment of functional impairment. In conjunction with the results of other tests, an attempt is made to estimate the degree of handicap and predict the need for and effects of rehabilitative programs, including the selection and fitting of hearing aids. Although speech recognition testing provides useful information for the benefit of the patient, it appears to be a lesser contributor than the pure-tone audiogram in the fitting of hearing aids. There are many models for selecting hearing aids on the basis of pure-tone threshold information only, but none for selecting hearing aids by relying solely on the results of speech recognition testing and ignoring the audiogram.

The important question to be asked is, Why does one bother to use speech recognition testing at all? Generally, tests are developed using stimuli which are easily quantified and can be varied along one dimension at a time. An effort is made to simplify patient judgments, patient responses, and scoring. Confounding variables are avoided. Speech recognition testing uses stimuli that are poorly quantified and controlled, that typically vary simultaneously along several dimensions, that involve relatively high-level patient judgments and responses, and that often require interpretive scoring. Results are confounded by linguistic, intellectual, and personality variables. Clearly, the development of speech testing did not proceed from an analysis of end-organ function or any model of how disease affects the way it works.

Why has speech recognition testing been pursued in the clinic? A large part of the answer comes from a frustrated wish to assess social adequacy. If that could be done, then it might be possible to quantify the impairment of hearing loss and evaluate the effectiveness of rehabilitative procedures. Clearly, speech communication is one of the important factors of social adequacy which is affected by hearing loss, giving it a strong face validity as a correlate of magnitude of handicap. This is so compelling that clinicians sometimes behave as if speech recognition were the only aspect of impairment important to patients. Audiologists select hearing aids for maximum intelligibility rather than for the most pleasing perception of music or voices. They tell patients that they must adapt when they complain about harshness. Patients must be made to understand that speech intelligibility is the most important priority. In the clinic, there is a tendency to behave as if speech recognition were social adequacy.

Assuming that speech communication is the major concern in assessing impairment, and that by knowing a patient's speech recognition ability his/her life can be improved, it does not necessarily follow that direct measurement of speech recognition will be the most satisfactory method of determining the patient's abilities. For one thing, an adequate sample of all the variations of speech the patient will encounter is impossible to assemble. Nor could appropriate weightings for frequency of occurrence or importance be given. As compromise is made for the creation of a practical test, face validity becomes less apparent; in its place one is obligated to define how well the sampling of speech recognition abilities predicts the patient's over-all speech communication performance.

Once it is realized that one is predicting or estimating speech reception capability rather than directly measuring it, the question should be asked about whether there might be better ways to accomplish that goal than direct measurement of selected speech recognition tasks. One does not measure a pure-tone audiogram in order to learn how a patient will hear pure-tone signals in his/her environment. This simple probe signal is used to describe altered function of the ear, and accurate predictions can often be made regarding the perception of complex stimuli. It may be interesting to note that the pattern of pure-tone thresholds may correlate as high as 0.70–0.80 with clinically measured speech recognition scores. Decreased speech recognition ability is a function of the damage which has occurred to the auditory system. It follows that if how speech is perceived and how function was altered in an impaired ear could be defined, then one could simply measure the changes in function and predict the effects on speech perception. To do this, more aspects of function than threshold sensitivity, loudness, and adaptation must be examined.

It cannot be denied that speech recognition testing has been used to advantage for clinical studies of hearing loss. The assumption that the goal of improving the assessment of speech reception abilities in the hearing impaired is best met by trying to improve speech recognition testing per se can be questioned. There is a need for continued work on understanding all the parameters of speech perception, particularly with impaired ears, plus a fundamental need to understand different manifestations of hearing loss. There is an incredibly weak understanding of auditory dysfunction. Witness the number of studies which find it sufficient to group data according to conductive and sensorineural hearing loss. Even the studies that attempt to match audiograms ignore the wide variation in pathologies that may give the same threshold configuration. Unless an attempt is made to understand the mechanisms which cause one ear to hear speech differently from another, real progress cannot be made. Perhaps some of the problems might be related to satisfaction with the definition of hearing loss in terms of pure-tone threshold sensitivity and speech recognition score.

Shortly before his death, Werner Heisenberg (1976) published an article in *Physics Today* in which he pointed out that when physicists begin with poor philosophy, they may pose the wrong questions. "It is only a slight exaggeration to say that good physics has at times been spoiled by poor philosophy" (p. 32). This might serve as a reminder to examine the question whenever the answer never seems to be satisfactory.

It is doubtful that speech recognition testing will soon be replaced, and it would appear that the current objectives of testing might be better met by improved tests. For those who may develop these tests, it will be important to keep in mind that many patients cannot read well or write legibly; they may not speak English and often may have difficulty with simple instructions; and they have widely varying levels of attention and motivation. Cumbersome administration and complicated scoring will lead to low acceptance by audiologists.

Finally, the error measurement should be clearly defined. What is meant by "high reliability" and what exactly does that mean to the clinician? Two recent articles (Dubno & Dirks, 1982; Dubno, Dirks, & Langhofer, 1982) point out that high reliability is seen across equivalent sets of 10 or 11 items. Interlist variance could be shown to be very small. However, the error of measurement for single administrations on a single subject was never shown. This is absolutely essential for test interpretation. Despite repeated claims of high reliability, the measurement error shown for sets of 100 items differed little from that predicted by simple binomial sampling theory and was no better than that for other tests of the same length.

It should be remembered that the clinician is making repeated tests on the same patient and must have a way of knowing whether or not observed test scores differ significantly from one another. The developer and certainly the promoter of any test must accurately provide that information in a usable form. It should not be couched in terminology to make the test look good and it should recognize that clinicians are typically not good statisticians. Again, error of measurement and confidence limits can assist the clinician. Reliability coefficients may be used to sell a test.

It should also be remembered that tests scored in terms of percent correct often have binomial error distributions despite the best efforts and intentions to avoid this unpleasant circumstance. The nature of error distributions should be examined very carefully and uniform variance should not be presumed. Likewise, standard deviations may not be very useful with skewed distributions and confidence limits may not be uniform across test scores.

One of the reasons it is so difficult to reduce interlist variance below that predicted by random sampling is that it is not reasonable to expect lists to be equated for a heterogeneous population such as that encountered with hearing loss. When different subjects are utilizing different cues, lists equated for one of the subjects would place a heavy weighting on those specific cues. That, of course, would do little for the second patient. In practice, equivalent lists have the same general characteristics and the average performance across subjects is equalized. A set of lists could be tailored for any given patient and measurement error could be reduced below binomial predictions, but it would not be feasible for tests which will be used with general clinical populations.

In summary, two basic points have been discussed. First, is the proper question being asked about the future direction of speech recognition testing? Second, if new tests are developed, they should be practical and well defined in terms of measurement error.

REFERENCES

DUBNO, J. R., & DIRKS, D. D. (1982). Evaluation of hearing-impaired listeners using a Nonsense Syllable Test. I. Test reliability. *Journal of Speech and Hearing Research*, *25*, 135–141.

DUBNO, J. R., DIRKS, D. D., & LANGHOFER, L. R. (1982). Evaluation of hearing-impaired listeners using a Nonsense Syllable Test. II. Syllable recognition and consonant confusion patterns. *Journal of Speech and Hearing Research*, *25*, 141–148.

HEISENBERG, W. (1976). The nature of elementary particles. *Physics Today*, *29*(3), 32–38.

Chapter 6

ARTICULATION TESTING METHODS FOR EVALUATING SPEECH RECEPTION BY IMPAIRED LISTENERS

LOUIS D. BRAIDA

Massachusetts Institute of Technology
Cambridge

An important clinical application of speech reception testing is the determination of appropriate characteristics for hearing aids for individuals with various types of hearing loss. Although many clinical procedures used in hearing aid fitting employ some form of speech reception tests, their role in the selection process is necessarily limited by the relatively high cost of discriminating among a large number of candidate instruments by experimental means. A more economical use of clinical testing time would be directed at determining the values of a set of patient variables which, within an appropriate theoretical framework, could be used to determine at least certain parameters of the aid most appropriate for the patient. In this context speech reception tests would continue to play two important roles. First, laboratory speech tests would be used extensively in the development of the theory serving as the basis for aid selection. Second, clinical speech tests would be used to evaluate the selection process and possibly to choose from among a small number of alternative aids. This paper focuses on the first of these two applications.

An earlier report (Dugal, Braida, & Durlach, 1980) outlined an approach to the problem of determining the optimum frequency-gain characteristics for hearing aids based on the use of Articulation Theory (Fletcher & Galt, 1950; French & Steinberg, 1947; Kryter, 1962). This work is part of an effort to develop a theoretical model capable of predicting the dependence of speech reception performance on the characteristics of the amplification system, the acoustic environment in which the hearing aid is used, and the properties of the hearing impairment. Based on a rather crude model of the hearing loss, we were able to use Articulation Theory to make relatively accurate predictions of the shapes of the performance-intensity (PI) functions for the impaired listeners studied by Skinner (1980) and the dependence of these functions on the frequency-gain characteristic. Further, we were able to use the theory to derive theoretical predictions for the "optimum" characteristics for these listeners which were in relatively good agreement with those derived from the empirical speech test results.

To the extent that these results can be generalized, the role of clinical speech reception testing in hearing aid fitting may be considerably modified because the range of characteristics that need be considered for a given listener can be greatly restricted on theoretical grounds. However, the extent to which Articulation Theory can provide the basis for such a model is presently unknown. The listeners for whom predictions were made were relatively homogeneous in terms of both etiology (noise-induced loss) and audiogram shape (high-frequency loss above roughly 1–2 kHz) and differed primarily with respect to severity of loss. In addition, the range of characteristics for which theoretical predictions could be checked included variation in the degree of high-frequency emphasis relative to a response which "mirrored" the audiogram. Further, there are some indications (DeGennaro, 1978) that the characterization of the impairment used, an equivalent masking noise, may be more suitable for sloping high-frequency losses than for flat losses.

In order to provide a firmer basis for determining the applicability of Articulation Theory, it is necessary as a first step to test its predictions both for a wider range of listeners and for a wider range of frequency-gain characteristics than that used in preliminary work. Although there have been many studies of the effects of varying the frequency-gain characteristic of amplification systems or hearing aids on speech reception for listeners with hearing impairments (Braida et al., 1979), these are generally inadquate for this purpose. Most previous studies of speech reception by impaired listeners have been critically deficient with respect to stimulus specification and control, speaker and listener training, and systematic variation in frequency-gain characteristic and presentation level to permit careful evaluation of the theory. In the balance of this paper, considerations relevant to the design of a study more suitable for this goal are discussed. The next section reviews the problem from the historical perspective relevant to the formulation of Articulation Theory as a means of predicting the performance of communication systems. The third section discusses the adaptations of these approaches required by the specific application to individual listeners with hearing impairments. The last section discusses a recent study of speech perception by impaired listeners which employed this method.

BACKGROUND

Articulation Theory was developed by communication engineers to facilitate the evaluation of communication systems by minimizing the need for speech reception testing of individual systems. The theory seems to have evolved from the realization that (a) the results of a wide variety of such tests conducted to evaluate the effects of varying the frequency response of the systems, (b) the addition of background noise, and (c) the presence of certain nonlinear distortions could be unified within a relatively simple theoretical framework. Underlying this approach were a set of commonly held assumptions about the nature of the speech communication process which bear review because their ultimate justification has relied upon the success of the approach as a whole rather than on detailed evaluation. It will be evident that certain of these assumptions are not satisfied in many clinical situations, and this raises difficulties for the straightforward application of the communication theoretic approach to clinical problems. In the third section we argue that with appropriate adaptations many of these limitations can be overcome.

A fundamental assumption of the approach adopted by communication engineers is the belief that the problem of speech communication can be separated into (a) the engineering problem of providing adequate transmission of the acoustic speech signal and (b) the psychological problem of interpreting the received signal to deduce the message intended by the speaker. Within this context it is impossible for the engineer to know, in an absolute sense, how well a given transmission system will function, since this depends on factors that are independent of the acoustic properties of the received signal, such as the size of the message set, the clarity with which the messages were enunciated, the competence of the listener in utilizing contextual cues, and so on. Engineering attention focused rather on predicting the relative performance of transmission systems under conditions in which such factors are kept constant.

The design of speech reception tests to be used in the comparative evaluation of communication systems has stressed the use of highly trained crews of speakers and listeners, with the implicit assumption that the substitution of an alternate speaker or listener for a member of the test crew would not change the test results. The use of highly trained crews facilitates the use of test materials such as nonsense syllables which sample adequately the important speech sounds of the language and whose reception does not depend on the extraneous contextual cues. Subsequently, a procedure based on a "proficiency factor" was developed to account for differences in the training of listeners and in the clarity of speakers. Techniques were developed also that permitted comparison of results obtained with different types of speech materials (e.g., nonsense syllables, monosyllabic words, sentences) and different message set sizes. Ultimately, the structure of Articulation Theory evolved so that the properties of the communication system (including the presentation level) determined a measure (the Articulation Index) of the relative audibility of signal components important for speech reception, the proficiency factor accounted for the ability of a given speaker-listener pair to utilize the system, and empirical functions were then used to estimate scores on various speech reception tests. It is noteworthy that the computation of the Articulation Index integrates properties of the transmission system (e.g., frequency-gain characteristic), the listener (absolute thresholds, spread of masking, discomfort levels), and the talker (speech spectrum and level distribution), as well as the general characteristics of speech (e.g., the band importance weights).

The speech reception testing approach adopted by communication engineers largely reflects this theoretical framework. It is important to consider the extent to which it is also representative of situations in which the communication systems that were studied are likely to be used. For example, the decision to employ trained speaking and listening crews appears eminently sensible if it is assumed that the communication situation being modeled is one in which the talker and the listener are familiar with the properties of the transmission system. The decision to focus tests on speech syllables is based on the long held belief that correct reception of syllable-sized segments is sufficient (if not necessary) for correct reception of messages composed of larger segments. Although attempts to interpret results on tests using polysyllabic English words have generally been consistent with this assumption (Hirsh, Reynolds, & Joseph, 1954), some studies concerned with reception of connected discourse (Giolas, 1966; Giolas & Epstein, 1963; Speaks, 1967) or nonsense sentences (DeGennaro, Braida, & Durlach, 1981) have been interpreted more negatively. It is presently unclear how to weight the latter results since only very simple linear relations between word intelligibility and discourse reception have been considered. Further, everyday experience with communication systems designed in accordance with Articulation Theory (e.g., the telephone system) attests to the sufficiency of the above assumption.

The attractiveness of the communication theoretic approach to the analysis of systems used by normal-hearing listeners for speech transmission ultimately stems from its effectiveness in predicting the performance of such systems. In this connection it is noteworthy that in many cases, particularly those involving the alteration of speech transmission characteristics by linear filtering or additive noise, Articulation Theory can predict the relative performance of different systems with an accuracy comparable to the reliability typically achieved in speech reception tests. Further, the fundamental aspects of the theory have survived rigorous experimental test.

APPLICATION TO IMPAIRED LISTENERS

In attempting to develop a form of Articulation Theory suitable for hearing-impaired listeners, investigators have adapted the communication theoretic approach used for speech transmission systems. As noted above, the fundamental assumption involves a separation of variables into those that affect the reception of the acoustic elements of speech and those concerned with the interpretation of the elements received. Within the framework of the theory, the properties of a given listener affect both types of variables, because (a) the state of auditory function is involved in the determination of the Articulation Index and (b) the listener's ability to utilize the speech cues provided by his/her auditory system affects the Proficiency Factor. It is also possible that the relationship between the Articulation Index and test scores is different than for normals. Consequently, in evaluating the usefulness of Articulation Theory for impaired listeners it is important to employ a strategy and testing methods that allow these possibilities to be analyzed separately. To distinguish between these possibilities, it is important to characterize both the test materials and the listeners who participate in the speech reception tests.

Two properties of the speech materials must be established: the distribution of speech levels in the frequency bands used to compute the Articulation Index and the function relating the Articulation Index to the predicted test score. Although representative values of speech level distributions have been published and are often used in Articulation Index calculation for normal-hearing listeners, these may not be adequate for hearing-impaired listeners who typically exhibit reduced dynamic range. The availability of speech analysis systems allows the level distributions and spectra of the actual test items to be determined (DeGennaro et al., 1981) and thus increases the accuracy of the Articulation Index calculation. Similarly, the function relating the Articulation Index and the predicted intelligibility score should be determined for the specific materials used. Since this function is assumed to be the same for all listeners, it can be conveniently estimated by testing listeners with normal hearing under conditions that lead to the same range of Articulation Index values as those obtained from the impaired listeners.

Within the context of Articulation Theory, the listeners must be described in terms of detection and discomfort thresholds over the range of frequencies that make significant contributions to the speech intelligibility. Measurement of these parameters is subject to both systematic and random errors which should be minimized to achieve good predictions for the Articulation Index. Two types of systematic errors must be controlled. First, differences in the specification of speech and hearing levels must be minimized, for example, by calibrating both under identical conditions in an acoustic coupler, or by measuring speech levels in a free field and using appropriate corrections for the transmission characteristics of the sound delivery system used. Second, detection and discomfort levels must be determined in a manner consistent with their role in the Articulation Index calculation procedure. Because measured values of these levels are known to be influenced by such experimental variables as instructions, it is important that they be validated relative to the actual speech materials. Unfortunately, techniques for accomplishing this have yet to be developed and constitute an important area for future research. To reduce the effects of random errors, such as fluctuations in measured hearing thresholds, it is important to repeat measurements throughout the period during which speech tests are administered.

An important methodological difference associated with the application to listeners with hearing loss stems from a lack of sufficient understanding of auditory function in cases of sensorineural impairment. Unlike the case of listeners with normal hearing, it is not presently possible to equate different impaired listeners and thereby construct the equivalent of listening crews. Thus, each impaired listener must be treated separately and regarded as his/her own control in testing the theory. Since the relation between the Articulation Index and the speech test score is generally expected to be nonlinear, systematic errors may be introduced if results obtained from different impaired listeners tested on identical stimulus conditions are averaged. As a result, it is necessary to test each impaired listener much more extensively than normal-hearing listeners. This dictates use of speech test procedures which can be used repeatedly with small long-term learning effects so that different presentation conditions can be compared reliably. For similar reasons, testing must be restricted to impairments that are relatively stable in time. In addition, it is probably unwise to assume that impaired listeners adapt to different listening conditions as readily as listeners with normal hearing, because many impaired listeners have extensive practice with the specific listening conditions provided by hearing aids which may not present adequately all the acoustic cues they are capable of processing. This requires that the listener be provided adequate training on materials processed by each presentation condition tested.

The need for materials that can be used repeatedly in a number of test conditions with small long-term learning effects can be satisfied by appropriately constructed nonsense syllable tests. For example, shorter learning effects have been observed for CV and VC items than for CVC items (Fletcher & Steinberg, 1929; Studebaker & Pavlovic, 1983), although performance on both shows similar dependence on test condition (presentation level, filtering, or *S/N* ratio). An efficient method for training listeners on closed set speech tests has been developed at C.I.D. (Miller, Engebretson, Garfield, & Scott, 1975). This method is particularly convenient if the test items are amenable to random access under computer control, so that feedback can be provided on a trial-by-trial basis and presentation probabilities can be altered dynamically to concentrate practice on the more difficult distinctions. However, the combination of small sets of speech materials and intensive training facilitates the learning of acoustic artifacts of no phonetic value, and this can make test results very difficult to interpret. This problem can be avoided if large numbers of tokens are available for each stimulus type (e.g., multiple utterances produced by multiple speakers) or if separate sets of tokens are used for training and testing. The use of tokens produced by more than one speaker should also reduce the dependence of test results on the clarity with which a given speaker enunciates the test items (Chen, Zue, Picheny, Durlach, & Braida, 1980; Picheny, Durlach, & Braida, 1980).

Another consideration relevant to the specification of speech test materials is the desired relation between the Articulation Index and the test score. To minimize the effects of random errors in the calculation of the Articulation Index, it is desirable that this function not be locally steep, but rather grow at a relatively uniform rate. In general, the steepness of this function depends on the number of stimulus-response alternatives available to the listener. Tests with small numbers of alternatives (e.g., digits) have functions much steeper than those for larger numbers of alternatives (e.g., 1,000 monosyllabic words). However, large stimulus sets generally require correspondingly large amounts of listener training before asymptotic performance is reached. It is therefore necessary to strike a balance between these two conflicting requirements.

EXPERIMENTAL STUDY

The appendix to this paper (by Milner, Braida, Durlach, & Levitt, 1983) reports on a study (Milner, 1982) of the effect of filtering on consonant identification by impaired listeners that was designed to provide data with which to test the predictions of Articulation Theory for these listeners. Six impaired ears and three normal ears were tested on 10 filtering conditions at a variety of presentation levels for each condition. To help evaluate the assumption that the hearing impairments can be adequately described in terms of elevated detection thresholds and reduced dynamic range, the listeners with normal hearing were also tested in the presence of masking noise shaped to produce the detection thresholds of certain of the impaired listeners. The results of these tests also provided data for determining the shape of the function relating the Articulation Index to predicted intelligibility score for the test materials.

Natural speech (CV nonsense syllables) was sampled and digitized to provide test items for this study. Measurements of ⅓-octave spectra and band-level distributions for these materials indicate a somewhat wider range (40–50 dB) of levels present during speech in most frequency bands, even after normalization for overall long-term level, than is commonly reported (Dunn & White, 1940). These measurements also indicate considerably greater dependence of the shape of the band-level distributions on frequency: Distributions are bimodal at low frequencies and become more highly peaked at higher frequencies. In addition, the long-term spectrum of these speech materials exhibits less roll-off above 1000 Hz than Dunn and White reported, but it is consistent with more recent data (Byrne, 1976).

Three tokens of each syllable were recorded by each of four speakers, with two of these tokens used for training and the third reserved for testing. The identity of the token used for testing changed from condition to condition. The filtering required in the different test conditions was achieved digitally using techniques that introduced minimal phase distortion, since such degradation was not the focus of this study. Speech tests were administered individually under computer control. This allowed the listeners to proceed at their own pace and facilitated subsequent data analysis. Each CV syllable was tested once for each speaker (576 items per condition), although selected conditions were retested to check the stability of the results and to obtain statistically reliable data on the pattern of errors made under the various processing conditions.

Speech reception results were described in terms of performance-intensity functions which relate test score to presentation level. For the normal-hearing listeners, performance-intensity functions generally exhibited the expected properties: Scores increase with level until a plateau is reached and at any given level, scores decreased when spectral information was removed. However, maximum performance for the high-pass 700-Hz condition was roughly equal to maximum performance for the unfiltered condition. In order of diminishing maximum scores, the remaining conditions were ordered as follows: low-pass 2800 Hz, band-pass 700–2800 Hz, high-pass 1400 Hz, band-pass 1400–2800 Hz, low-pass 1400 Hz, band-pass 700–1400 Hz, low-pass 700 Hz and high-pass 2800 Hz. Relative to the results reported by others for normal-hearing listeners (French & Steinberg, 1947), these data indicate superior performance for low-pass 700 Hz, low-pass 1400 Hz, and high-pass 2800 Hz; slightly inferior performance for high-pass 700 Hz and high-pass 1400 Hz; and roughly equivalent performance for the wideband and low-pass 2800-Hz conditions.

The listeners with sensorineural hearing losses generally exhibited a dependence of performance on filter condition that was similar to the normal-hearing listeners in that removing spectral information generally reduced intelligibility at a given presentation level. Maximum scores in each condition, including the unfiltered speech, were lower than for normal-hearing listeners and also depended on the degree of hearing losses. The specific contribution of the highest frequency band (high-pass 2800 Hz) did not appear to be as large for these listeners since scores did not increase greatly when this band was added to conditions containing information below this cut-off frequency. However, at the higher presentation levels, removing information below 700 Hz resulted in maximum scores equal to or better than the scores for those conditions in which this information was present. Finally, the performance-intensity functions for these listeners exhibited significant roll-over, at the highest levels tested, much more frequently than for the normals. The listeners with simulated losses exhibited higher performance than the listeners with real losses, although the dependence on filtering condition was roughly the same. Generally, these listeners exhibited roll-over for the same test conditions (700 Hz low-pass, 700–1400 and 1400–2800 Hz band-pass) as the listeners with sensorineural hearing loss.

Subsequent reports will discuss these results from the point of view of Articulation Theory and analyze the patterns of errors on the speech tests.

REFERENCES

Braida, L. D., Durlach, N. I., Lippman, R. P., Hicks, B. L., Rabinowitz, W. M., & Reed, C. M. (1979). *Hearing aids: A review of past research on linear amplification, amplitude compression, and frequency lowering* (ASHA Monographs Number 19). Rockville, MD: American Speech-Language-Hearing Association.

Byrne, D. J. (1976). The speech spectrum—Some aspects of its significance for hearing aid selection and evaluation. *British Journal of Audiology, 11*, 40–46.

Chen, F. R., Zue, V. W., Picheny, M. A., Durlach, N. I., & Braida, L. D. (1980). Speaking clearly: Acoustic characteristics and intelligibility of stop consonants. *Journal of the Acoustical Society of America*, 67(Suppl. 1), S38.

DeGennaro, S. V. (1978). *The effect of syllabic compression on speech intelligibility for normal listeners with simulated sensorineural hearing loss.* Unpublished master's thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA.

DeGennaro S. V., Braida, L. D., & Durlach, N. I. (1981). Statistical analysis of third–octave speech amplitude distributions. *Journal of the Acoustical Society of America*, 69(Suppl. 1), S16.

Dugal, R. L., Braida, L. D., & Durlach, N. I. (1980). Implications of previous research for the selection of frequency-gain characteristics. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance and measurement* (pp. 379–403). Baltimore: University Park Press.

Dunn, H. K., & White, S. D. (1940). Statistical measurements on conversational speech. *Journal of the Acoustical Society of America, 11*, 278–288.

Fletcher, H., & Galt, R. H. (1950). Perception of speech and its relation to telephony. *Journal of the Acoustical Society of America, 22*, 89–151.

Fletcher, H., & Steinberg, J. C. (1929). Articulation testing methods. *Bell System Technical Journal, 8*, 806–854.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19,* 90–119.

Giolas, T. (1966). Comparative intelligibility scores of sentence lists and continuous discourse. *Journal of Auditory Research, 6,* 31–38.

Giolas, T. G., & Epstein, A. (1963). Comparative intelligibility of word lists and continuous discourse. *Journal of Speech and Hearing Research, 6,* 349–358.

Hirsh, I., Reynolds, E. G., & Joseph, M. (1954). Intelligibility of different speech materials. *Journal of the Acoustical Society of America, 26,* 530–538.

Kryter, K. D. (1962). Methods for the calculation and use of the Articulation Index. *Journal of the Acoustical Society of America, 34,* 1689–1697.

Miller, J. D., Engebretson, A. M., Garfield, S. A., & Scott, B. L. (1975). New approach to speech-reception testing. *Journal of the Acoustical Society of America,* 57(Suppl. 1), S48.

Milner, P. (1982). *Perception of filtered speech by hearing-impaired listeners and by normal listeners with simulated hearing loss.* Unpublished doctoral dissertation, City University of New York.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1980). Speaking clearly: Intelligibility and acoustic characteristics of sentences. *Journal of the Acoustical Society of America,* 67(Suppl. 1), S38.

Skinner, M. W. (1980). Speech intelligibility in noise-induced hearing loss: Effects of high-frequency compensation. *Journal of the Acoustical Society of America, 67,* 306–317.

Speaks, C. (1967). Intelligibility of filtered synthetic sentences. *Journal of Speech and Hearing Research, 10,* 289–298.

Studebaker, G. A., & Pavlovic, C. V. (1983). A nonsense syllable test designed for articulation index testing. *Journal of the Acoustical Society of America,* 73(Suppl. 1), S102.

APPENDIX

PERCEPTION OF FILTERED SPEECH BY HEARING-IMPAIRED LISTENERS

Paul Milner
Louis D. Braida
Nathaniel I. Durlach

Massachusetts Institute of Technology, Cambridge

Harry Levitt

City University of New York, New York

Six normal-hearing and hearing-impaired subjects listened to 72 consonant-vowel (CV) nonsense syllables spectrally limited by 10 conditions of low-pass, high-pass and band-pass filtered speech. The syllables were heard at a minimum of five different presentation levels ranging from near threshold to near discomfort level. All listeners heard the speech samples with no added background noise. In addition, the normal-hearing subjects listened to the speech materials in a noise background spectrally shaped to simulate the hearing loss of selected hearing-impaired subjects. The results of this study have shown that for all subjects, intelligibility of the materials used remained relatively high whether high-pass filtered at 700 Hz or low-pass filtered at 2800 Hz, when heard with sufficient intensity. The performance-intensity functions of the normal-hearing listeners with simulated high-frequency hearing loss were similar to those of the listeners with relatively flat sensorineural hearing loss. Roll-over in performance occurred at high presentation levels that were below reported discomfort thresholds.

INTRODUCTION

The purpose of this study is to attempt to determine how speech intelligibility varies as a function of both spectral content and intensity level for a group of listeners with different sensorineural hearing impairments. Identical tests were performed on normal-hearing listeners to determine baseline measures for the intelligibility tests. In addition, the normal-hearing listeners were tested with additive noise to simulate the hearing loss of certain of the impaired listeners. The present report describes and discusses the data collected for six normal-hearing and hearing-impaired subjects listening to 10 conditions of filtered speech over a wide range of intensities.

Subjects

Six subjects participated in this study. Three subjects had bilateral sensorineural hearing loss, one had a unilateral sensorineural loss, and two subjects had normal hearing. All subjects underwent complete otological and audiological evaluations prior to testing. None showed any significant otopathology other than that related to their hearing impairments.

In this study the listeners with normal hearing were also measured under masking conditions simulating hearing loss, created by introducing shaped wide-band noise masking in the test ear at fixed level. The intensity and shape of the masker were selected to produce masked thresholds similar to the audiograms of the subjects with impaired hearing. Table 1 lists the ⅓-octave band noise levels used to create the simulated loss for each subject.

Table 1. Masking noise for simulated hearing loss for three normal-hearing listeners.

Band center frequency	*Band pressure level (dB re 20 μPa in NBS-9a coupler)* J.G.	J.T.	R.M.
200	24	24	24
250	26	26	26
315	28	30	28
400	32	30	31
500	34	33	34
630	36	35	37
800	38	36	40
1000	43	38	43
1250	47.5	39.5	48
1600	51	43	53
2000	54	45	57
2500	57	48	59.5
3150	60	51.5	64
4000	61	53	67.5
5000	61	53	66
Overall level (C-weighted)	70	68	74

Table 2 indicates the audiometric data for each subject as well as the masked pure-tone thresholds for the normal-hearing subjects when tested with simulated hearing loss. Although masked thresholds were only obtained up to 4 kHz, the masking noise spectrum continued beyond 10 kHz and was limited only by the characteristic of the earphones used for the experiments (Telephonics Model TDH-49).

TABLE 2. Subject audiometric data and masked pure-tone thresholds for the normal-hearing subjects.

				Frequency in Hz				
Subject	*250*	*500*	*1K*	*2K*	*4K*	*8K*	*SRT*	*%W-22*[a]
J.T., Age 26, Female								
Right	0	0	0	0	0	5	0	100
R/Mask	–	20	30	35	50	–	–	–
R.M., Age 67, Female								
Right	15	10	15	15	15	5	10	100
R/Mask	–	25	30	45	55	–	–	–
J.G., Age 43, Female								
Etiology: Right ear only—sudden onset, unknown								
Left	5	5	5	5	10	15	5	100
Right	15	20	35	40	65	105	25	88
L/Mask	–	25	35	40	55	–	–	–
F.G., Age 48, Male								
Etiology: Noise exposure—both ears								
Left	5	5	15	35	95	90	15	96
Right	5	0	15	30	70	85	10	96
G.M., Age 70, Male								
Etiology: Presumed presbycusis—both ears								
Left	5	15	20	35	75	60	15	88
Right	20	25	45	40	50	85	35	80
T.T., Age 27, Female								
Etiology: Congenital (bilateral)								
Right	70	65	70	75	75	55	65	88

Note. All levels are in dB HL (re ANSI-1969).
[a]%W-22 = % discrimination using CID W-22 word lists tested at 40 dB SL (re SRT) except T.T., tested at 30 dB SL.

Speech Signals

Because consonant reception typically is more affected by hearing loss and filtering than vowel reception, the test materials were chosen to focus on consonant recognition. The speech signals for all experiments were 72 consonant-vowel (CV) syllables recorded in an anechoic environment by each of four different talkers—two male and two female. The 72 CVs consisted of all combinations of 24 initial consonants (p, t, k, b, d, g, f, th, s, sh, v, th, z, zh, ch, j, m, n, r, l, w, h, wh, y) and the three vowels /a/, /i/ and /u/. Each talker spoke each CV syllable three times, producing 864 utterances for the entire set of recorded CVs.

The recorded syllables were low-pass filtered at 4500 Hz and converted to 12-bit digital samples at a sampling rate of 10000 Hz. All the digitized speech waveforms were normalized to equal RMS (root-mean-square) levels relative to the vowel intensities.

In addition to the wide-band condition (actually 4500-Hz low-pass), nine other conditions of filtering were studied:

— Low-pass filtered at 700, 1400, and 2800 Hz
— High-pass filtered at 700, 1400, and 2800 Hz
— Band-pass filtered from 700–1400, 1400–2800
and 700–2800 Hz

Filtering was accomplished digitally using linear phase finite impulse response filters (McClellan, Parks, & Rabiner, 1973). All filters had a uniform transition region 200 Hz wide and a stop-band attenuation of at least 60 dB. While these filters had relatively sharp cut-off slopes and were free of phase distortion, out-of-band components were attentuated by 60 dB, unlike the case of analog filters which attenuate by progressive amounts outside the passband.

Experimental Set-Up and Procedures

Listening was done using TDH-49 earphones in a sound-proof room. Level control was accomplished using a Grason-Stadler Model 162 speech audiometer. Equivalent free-field (FF) sound pressure levels for the earphones were obtained using normal-hearing listeners to measure absolute auditory threshold differences between these earphones and a set of earphones (Telephonics Model 556—Villchur, 1970) previously calibrated to free-field sound pressure levels by Lippmann (1981). Table 3 shows the differences in thresholds between the two sets of earphones, the free-field correction from Lippmann, and the total correction applied to the Articulation Index calculation procedures (Kryter, 1962b) at each of the ⅓-octave band center frequencies.

TABLE 3. Free-field calibration data for the experimental procedure.

Frequency	*Threshold difference TDH-49 vs. 556*	*Free-field correction*[a]	*Overall correction*
200	−1.2	− 9.0	−10.2
250	+6.0	− 7.0	− 1.0
315	+1.8	− 4.0	− 2.2
400	+1.2	− 3.5	− 1.3
500	+1.4	− 6.5	− 5.1
630	+3.3	− 6.5	− 3.2
800	+3.4	− 3.5	− 0.1
1000	+1.8	− 6.0	− 4.2
1250	+2.2	− 5.0	− 2.8
1600	+1.3	− 4.0	− 2.7
2000	+5.9	− 2.5	+ 3.4
2500	+1.2	− 6.0	− 4.8
3150	−0.7	−10.0	−10.3
4000	−0.2	− 8.0	− 7.8
5000	−0.6	− 4.5	− 5.1

Note. All values in decibels.
[a]Lippmann, 1981.

Each subject was tested individually while seated in front of a computer video terminal in a sound-proof room. Two types of experimental procedures were used: a training procedure and an absolute identification procedure. For both procedures an utterance was selected at random from a predetermined list of utterances and played to the listener. The listener then typed into the computer one of 72 possible responses representing the syllable heard. A complete record of the experiment was contained in a computer data file for future analysis.

Subjects trained on the test materials and conditions by using immediate feedback when errors were committed (Miller, Engebretson, Garfield, & Scott, 1975). During a feedback interval the signal and response CVs were alternately played to the subject while the display indicated which utterance was being played.

In the identification experiments, the CVs were selected at random with equal probability from the entire set of utterances without replacement. No feedback was given and a new trial was initiated as soon as the response to a given trial was entered. If the subject did not respond, the utterance was returned to the list of unplayed tokens and selected again at random. To reduce the effect of hearing any token more than once, subjects were discouraged from skipping items unless a significant distraction occurred.

Presentation Conditions

Each subject heard all 10 filtered conditions at a minimum of five different presentation levels. Presentation levels usually extended over a 40–50-dB range, from about 20–30 dB below the estimated most comfortable loudness (MCL) for any condition to about 15–20 dB above. These limits were determined practically by a lower audibility limit and by discomfort at high presentation levels. Extra precautions were taken in F.G.'s case to keep presentation levels below the threshold of discomfort. This was done because he occasionally reported tinnitus after testing at relatively high intensity; however, he did not indicate that the levels were uncomfortably loud. If tinnitus was reported, no further testing in that ear took place on that day.

Initially, training experiments were conducted at the MCL determined for the specific filter condition. However, MCL was not used for any precise considerations in performance but rather was used as a guideline in determining the range of presentation levels to the subject. Subsequent training experiments, occurring after a series of identification experiments for the same filter condition, were run at or near the maximum performance level for the condition.

Contralateral masking was used in the nontest ear whenever any potential for cross hearing existed at higher presentation levels. The masking signal was the "speech noise" masker built into the Grason-Stadler speech audiometer. If the overall level of the filtered speech signals was sufficiently low, as in the high-pass 1400-Hz and high-pass 2000-Hz conditions, masking was reduced or eliminated to avoid the danger of overmasking.

RESULTS

Data for this study are plotted as performance-intensity functions for each filter condition. For these data, observed intelligibility scores were averaged across the four talkers and represent both consonant and/or vowel errors. That is, if either the consonant or vowel component of the CV utterance was incorrectly identified, the entire utterance was scored as incorrect.

Overall reliability of the individual scores was quite high. Direct estimates of test-retest reliability were performed on subjects for whom test points were repeated, and a split-half method of analysis was used for those conditions that were run only once. Generally, intelligibility differences of greater than 3 percentage points are significant.

Normal-Hearing Listeners

Performance versus Intensity (P-I) functions for the three normal-hearing listeners tested in quiet and with masking noise are shown in Figure 1. For the conditions tested in quiet, the functions exhibited characteristic performance improvements as presentation level increased and as the amount of available spectral energy increased. In noise, performance diminished at high presentation levels for several conditions.

For the unfiltered speech maximum performance was achieved at a presentation level of 62 dB SPL (FF) for all three subjects. This level is similar to that reported by French and Steinberg (1947). A more detailed comparison of the data of the present study with French and Steinberg's data is made later in this report.

Perfect performance was not achieved by any listener. One reason for this is the fact that the unfiltered speech has an upper limit of 4500 Hz. In addition, observed scores were influenced by the use of untrained talkers, the general difficulty of the speech materials (e.g., Hirsh, Reynolds, & Joseph, 1954), and the absence of a carrier phrase (Egan, 1948).

For the tests performed in quiet, each subject maintained a performance plateau over a 30–40-dB range in presentation level. For all filter conditions but one, maximum performance was achieved between 60 and 65 dB SPL. The exception was the HP (high-pass) 2800-Hz filter condition which, for subject J.T., showed increased intelligibility as presentation level increased. This is most likely due to the finite attenuation characteristics of the digital filters used. The filters were designed to have the stop band at 200 Hz above or below the cut-off frequency, with a stop-band attenuation of 60 dB. For the HP 2800-Hz filter, this meant that components below 2600 Hz were present but attenuated 60 dB relative to the pass band. At high relative presentation levels some of the more intense low-frequency vowel energy was probably audible, increasing intelligibility scores at those levels.

In comparing the results for R.M., an older adult, to the results of J.G. and J.T., one observes for R.M. somewhat steeper slopes of the performance-intensity functions and lower scores at the same presentation levels. The maximum scores were also slightly lower for all conditions, and in particular for HP 2800 Hz. This reduced performance may be due to slight hearing impairment in R.M. as well as possible age effects that are not well understood.

For the subjects tested with ipsilateral masking, P-I functions shifted as a function of intensity due to the threshold shift caused by the masking. Maximum scores achieved by each subject were generally lower than those of the unmasked conditions. Slopes of the functions were generally steeper, however, than for the conditions tested in quiet, especially for the high-pass and band-pass conditions.

Hearing-Impaired Listeners

Performance-intensity functions for the hearing-impaired listeners are presented in Figures 2 and 3. Subject F.G., with bilateral noise-induced high-frequency hearing loss, was tested in both ears because his left ear exhibited greater hearing loss above 2000 Hz than his right. Overall maximum performance for F.G. occurred for the HP 700-Hz condition in both ears, with little difference between their scores. However, maximum performance for the unfiltered condition in the left (poorer) ear was nearly 10 percentage points lower and was reached at 20-dB lower intensity than for the right. When energy below 700 Hz was removed, F.G. apparently was better able to use his residual high-frequency hearing. For the HP 2800-Hz condition, both ears exhibited maximum scores that were nearly equal to and at essentially the same presentation levels as for the normal-hearing listeners. As noted, due to the finite stop-band attenuation of the digital filters, some low-frequency energy was audible at high presentation levels. Since his hearing levels are nearly normal below 1000 Hz, that energy was, no doubt, audible to this listener.

It is also not surprising that performance at low-to-moderate levels for the LP (low-pass) conditions was comparable to that for the normals. It is likely, however, that at higher levels the intense low-frequency energy present for these conditions masked what little high-frequency energy may be available to this subject.

Observed performance relative to the low-pass conditions increased for the three band-pass (BP) conditions from the BP 700–1400 Hz to BP 1400–2800 Hz and BP 700–2800 Hz. Once again it is apparent that as low-frequency energy was removed from the speech heard by this subject, he was better able to utilize his residual high-frequency hearing capability.

For subject G.M. the difference, based on audiometric data, between his right and left ears was much greater than for F.G. This was also reflected in the different observed performance for each ear tested. In all cases, maximum scores in the left ear were higher and occurred at lower presentation levels than in the right.

In G.M.'s left ear significantly higher performance was obtained for the four widest band conditions: unfiltered speech, LP 2800 Hz, HP 700 Hz, and BP 700–2800 Hz. Similar to the results for F.G., the scores for the HP 1400-Hz and BP 700–1400-Hz conditions were lower and nearly equal. Intelligibility for the LP 700-Hz and HP 2800-Hz conditions were the poorest. Unlike for F.G., however, overall maximum performance in both ears was not achieved for the HP 700-Hz condition, but for the unfiltered speech. G.M.'s maximum scores were also lower overall than F.G.'s. Possibly due to factors associated with age and etiology, G.M. may have been less able to use residual high-frequency hearing, as F.G. seemed to be able to do.

As noted, scores in G.M.'s right ear were lower overall than in the left ear, even at high presentation levels. Highest scores again were achieved for the four widest band conditions. At the highest level tested, these scores were nearly equal. However, the slopes of the P-I curves appear significantly different. This effect was likely due to poorer low-frequency thresholds in the right ear, compared to the left. The somewhat better high-frequency thresholds in the right ear,

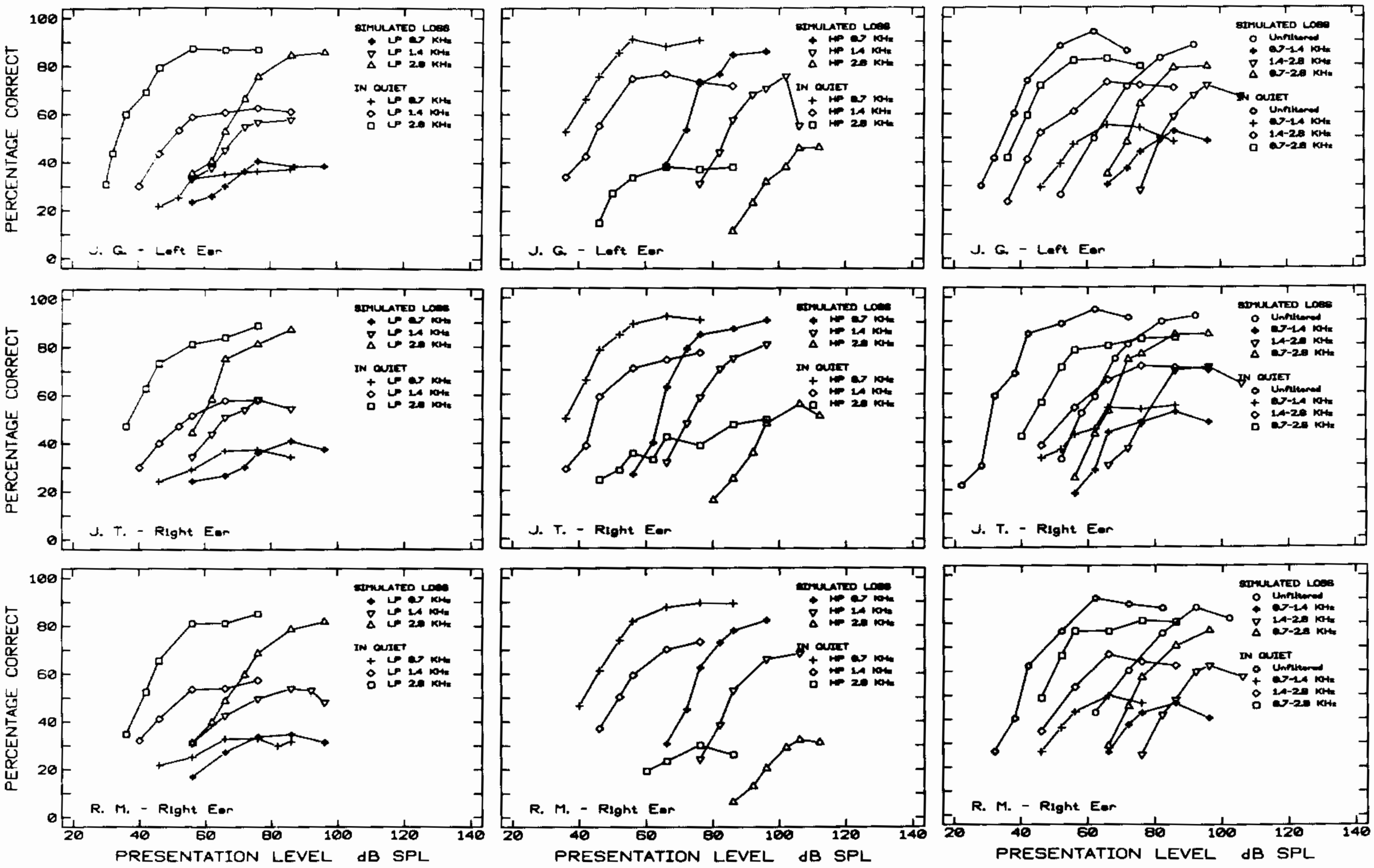

FIGURE 1. Performance-intensity functions for the listeners with normal hearing. LP = low-pass; HP = high-pass.

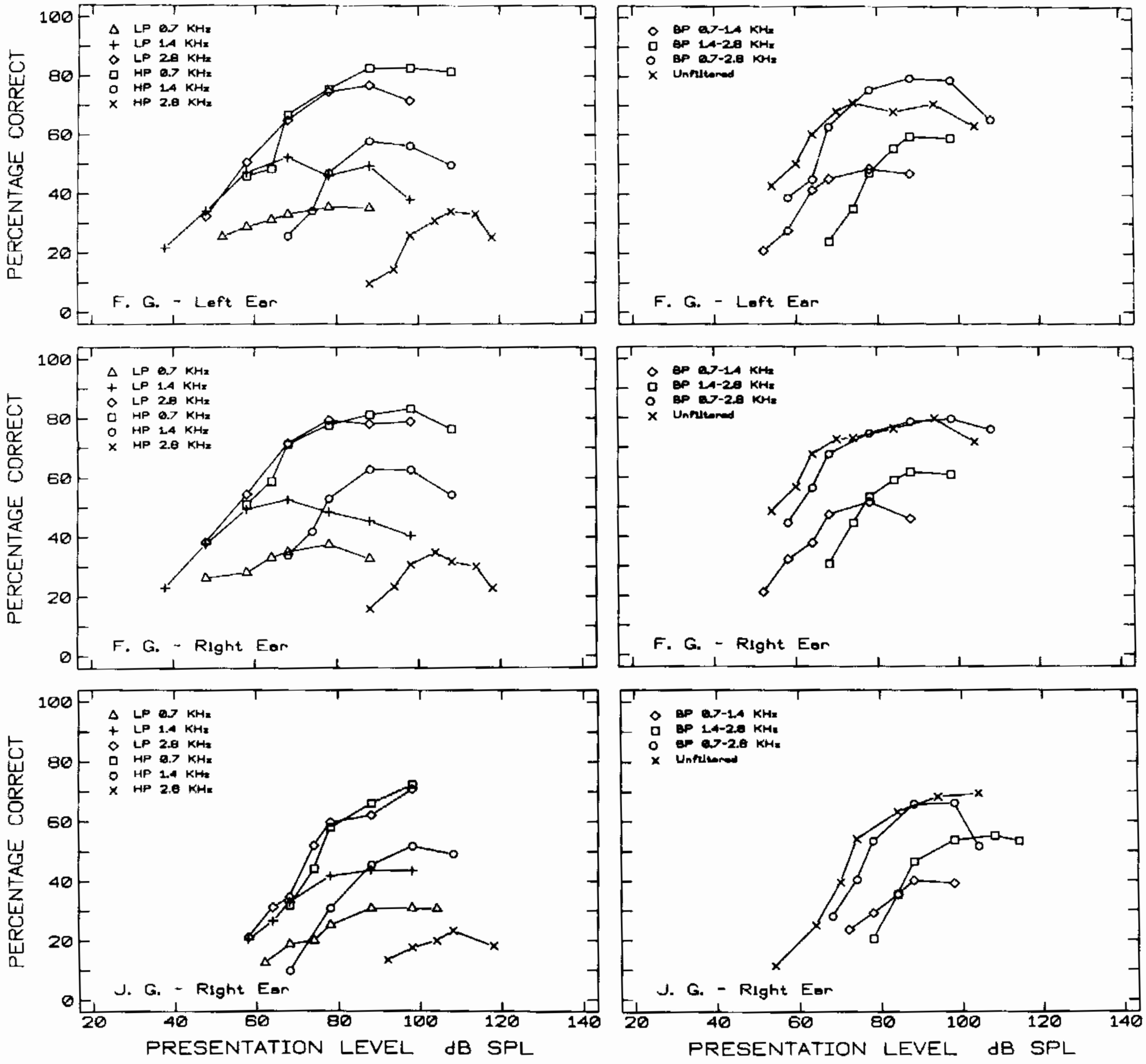

FIGURE 2. Performance-intensity functions for listeners F.G. and J.G., having impaired hearing.

however, allow his performance to improve as intensity increases, but not to the levels reached by the left ear. Although performance for the PB 700–1400-Hz and BP 1400–2800-Hz bands are nearly equal and relatively poor, performance for the resultant band, BP 700–2800 Hz, is substantially better.

Subject J.G. was tested in her impaired right ear. (J.G. was also tested in her left ear as a normal-hearing listener.) In comparing the 10 conditions, higher performance scores were obtained for the wider band conditions, as was true for G.M. and F.G. Observed performance for the high-pass 700-Hz condition at high presentation levels was greater than for the unfiltered condition. Poorest performance occurred for the HP 2800-Hz condition. The hearing loss in J.G.'s right ear was similar to that in G.M.'s left ear in shape and degree. Although J.G. is much younger than G.M., her maximum scores were generally lower than G.M.'s scores and occur at higher levels.

As for the other impaired listeners, subject T.T. achieved higher intelligibility scores under wide-band conditions than in the other filtered conditions. However, unlike the other impaired listeners, T.T. showed no improvement in the high-pass 700-Hz condition over the unfiltered condition at any presentation level. For T.T. the high-pass 2800-Hz band was barely audible and therefore contributed little useful information. The importance of the 1400–2800-Hz band appeared significant when this band was added to those below it in frequency.

Figure 4 compares the results obtained with the normal-hearing listeners tested in noise to the hearing-impaired subjects whose loss was closest in configuration to the simulated loss. In general, scores were higher for the simulated loss subjects than for the impaired subjects at similar intensities. The performance-intensity functions of the simulated hearing loss conditions also appeared to be more like those of the flatter loss subjects—J.G., T.T., and G.M.'s right ear—rather than the sloping high-frequency losses of F.G. or G.M.'s left ear.

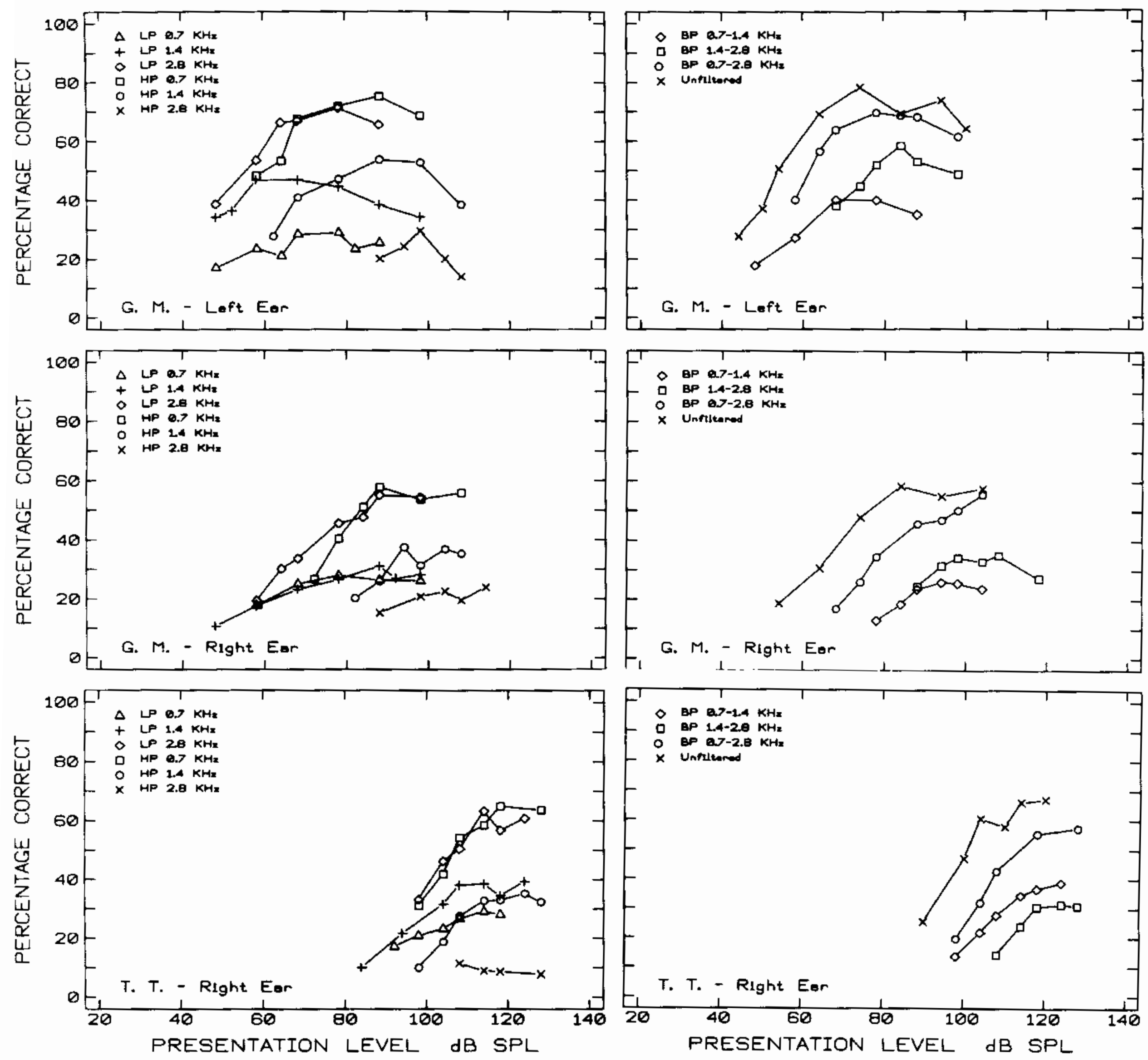

FIGURE 3. Performance-intensity functions for listeners G.M. and T.T., having impaired hearing.

DISCUSSION

For all subjects, intelligibility performance generally increased as presentation level increased and as more spectral components became available. However, the full spectrum was not always necessary for a subject to achieve maximum performance. This section discusses the results obtained when the lowest band (LP 700 Hz) or the highest band (HP 2800 Hz) or both simultaneously were filtered out. Cases in which performance was found to decrease (roll-over) as presentation level increased are examined. Finally, the results of several previous studies of the perception of filtered speech by normal-hearing listeners and by hearing-impaired listeners are compared to the results of the present study.

Influence of Spectral Components Below 700 Hz

In normal speech signals most of the energy is primarily in low frequencies. This region below 700 Hz in this study had some negative influences on intelligibility of speech at high presentation levels, both for normal-hearing listeners and for listeners with sensorineural high-frequency hearing loss. A comparison of the results for conditions of filtered speech with and without this energy for both normal-hearing and hearing-impaired listeners demonstrates these effects.

For the normal-hearing subjects, intelligibility of the HP 700-Hz condition at low presentation levels was lower than that of the unfiltered speech. However, at high presentation levels, intelligibility scores were equal or nearly equal to the unfiltered speech. This result was obtained when these listeners were tested with masking noise to simulate hearing loss.

For the impaired subjects the effect of removing energy below 700 Hz was dependent on the configuration of the hearing loss. For listener F.G., who had nearly normal hearing at low frequencies, the presence of this energy in unfiltered speech caused his performance to diminish when presented at normal and higher levels. When this

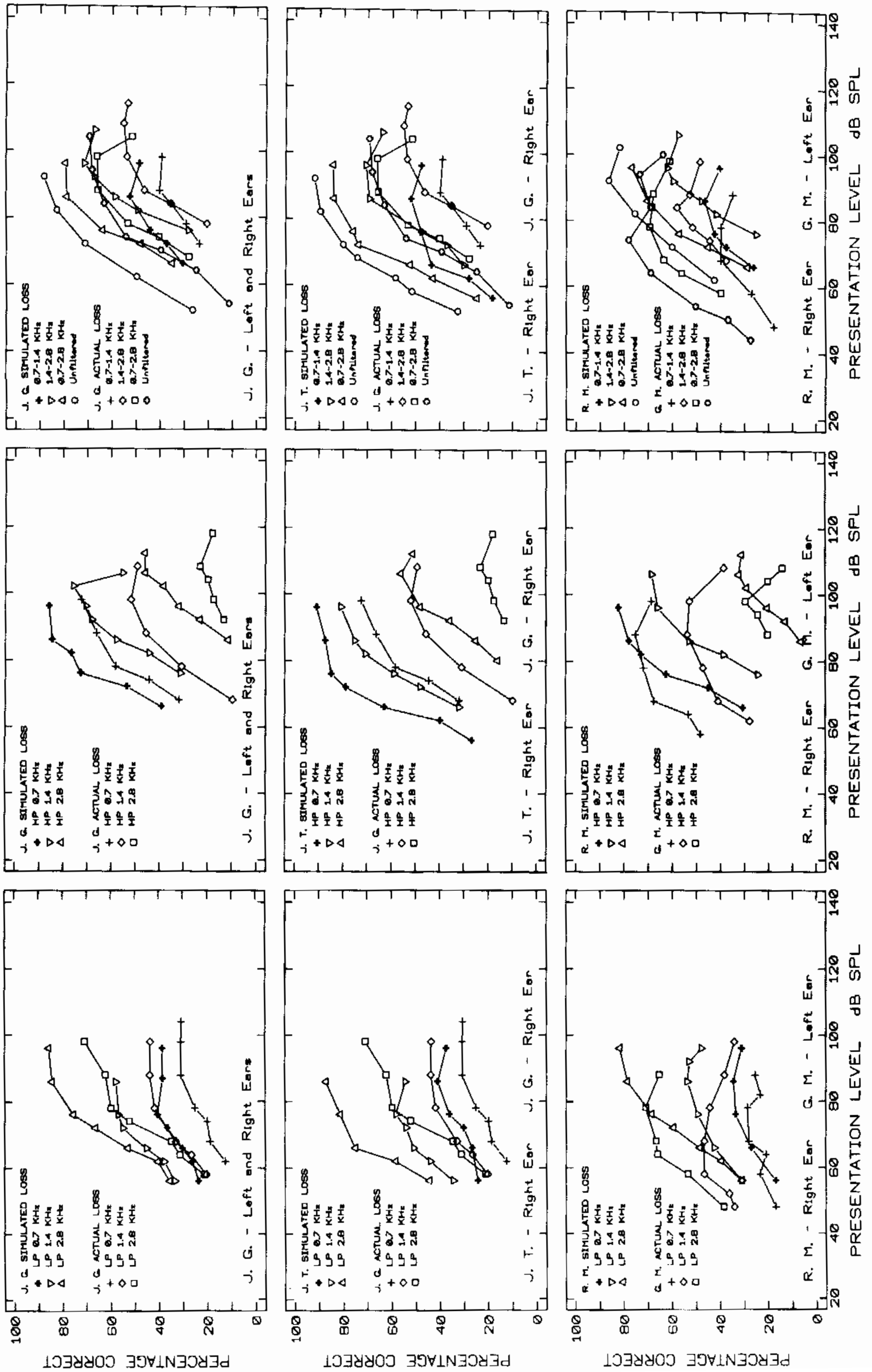

FIGURE 4. Comparison of performance-intensity functions for listeners with impaired hearing and for listeners with loss induced by masking noise.

energy was attenuated by filtering, scores increased at high presentation levels. The effect was greater in the left ear, which had a greater high-frequency hearing loss than the right ear. For the other impaired subjects only J.G. showed improved performance for the HP 700-Hz condition, but scores for the remaining subjects were generally no poorer than the unfiltered condition.

The influence of the energy below 700 Hz may also be seen in comparing intelligibility scores for the LP 1400-Hz and BP 700–1400-Hz conditions. For the normal-hearing listeners in quiet and in noise, intelligibility scores for the LP 1400-Hz condition were higher than for BP 700–1400-Hz at low presentation levels. However, at higher presentation levels, performance for BP 700–1400 Hz was nearly equal to LP 1400 Hz, indicating that energy at normal intensity levels below 700 Hz was not essential for relatively good speech intelligibility. It is also possible that as presentation levels increase for LP 1400 Hz, the upward spread of masking of low-frequency energy will limit any improvement in intelligibility.

Intelligibility scores nearly equal to the wide-band speech were found at high presentation levels in the HP 700-Hz and BP 700–2800-Hz conditions for measurements in J.G.'s impaired right ear and G.M.'s left ear, in which the loss was more steeply sloping in the high frequencies than was the loss in his right ear. In addition, for these listeners, when the LP 700-Hz band was added to BP 700–1400 Hz to produce LP 1400 Hz, intelligibility for the LP 1400-Hz filtered speech was higher than BP 700–1400 Hz at low intensities. As presentation level increased, differences between scores for LP 1400 Hz and BP 700–1400 Hz diminished and, at the highest levels tested, were nearly equal for each subject. For listener T.T. and the right ear of G.M., which have flatter hearing losses, the results were more like those of the normal listeners tested in quiet and in noise. In these listeners, removing low-frequency components did not result in the degree of improved performance relative to those conditions which contain these components, as it did for the normal-hearing listeners or listeners with sloping losses. However, scores for LP 1400 versus BP 700–1400 were much closer at high presentation levels, as they were for the sloping loss subjects just cited.

A possible cause of the effects observed for the LP 700-Hz band is the upward spread of masking of relatively intense low-frequency energy at high presentation levels (Bilger & Hirsh, 1956; French & Steinberg, 1947; Kryter, 1962a). This may reduce intelligibility for listeners with sloping high-frequency hearing loss. As seen in the results of F.G., significant improvements in performance resulted when energy below 700 Hz was removed by high-pass filtering, especially in his left ear which suffered from a somewhat greater high-frequency hearing loss than his right ear. It is also likely that because of the characteristics of the digital filters used in this study, part of the improvement in scores for the HP 700-Hz and BP 700–2800-Hz conditions at high levels resulted from some low-frequency energy being audible at very high presentation levels. However, the spread of masking would be reduced or eliminated because the low-frequency energy intensity is greatly reduced. Other researchers have shown that low-frequency energy at low intensities may contribute to intelligibility (e.g., Franklin, 1969, 1975; Rosenthal, Lang, & Levitt, 1975).

Influence of Spectral Components Above 2800 Hz

Whereas the low-frequency bands contribute more to the energy in speech, higher frequency bands contribute more to its intelligibility. The importance of components above 2800 Hz for the normal-hearing listeners is evident in Figure 1. Higher scores were achieved in the unfiltered speech and HP 700-Hz conditions than in the LP 2800-Hz and BP 700–2800-Hz conditions. When these listeners were tested in noise simulating the sloping hearing losses, the inclusion of the HP 2800-Hz band also improved intelligibility, but not to the degree observed for the tests in quiet. The results for HP 1400-Hz filtered speech—which, in effect, consists of adding HP 2800 Hz to BP 1400–2800 Hz—were also higher for the wider band condition for the tests both in quiet and in noise. For normal listener J.G., however, the improvement in quiet was not as great as for the other two normal listeners; and in noise, there was no observable difference in her scores for these two conditions.

For the hearing-impaired listeners (Figures 2 and 3) the importance of the HP 2800-Hz band was dependent on the subjects' hearing loss configuration. For the subjects with high-frequency sloping losses (F.G., J.G., and the left ear of G.M.) no differences were apparent in scores between HP 700 Hz and BP 700–2800 Hz. For the subjects whose losses were flatter (T.T., and G.M.'s right ear) adding the HP 2800-Hz band to BP 700–2800 Hz resulted in higher scores at low and moderate presentation levels. At the highest levels tested for these two subjects, however, there were virtually no differences between scores for any of the four conditions: unfiltered, LP 2800 Hz, HP 700 Hz, and PB 700–2800 Hz. In comparing HP 1400 Hz and BP 1400–2800 Hz, addition of the components above 2800 Hz made little or no difference in scores for any of the subjects except T.T., who achieved higher scores for the HP 1400-Hz condition than for the BP 1400–2800-Hz condition, particularly at the lower presentation levels. At the highest levels tested, however, the differences were smaller, although scores for the HP 1400 Hz were still higher.

Intelligibility scores for three of the four narrowest band conditions—LP 700 Hz, BP 700–1400 Hz, and HP 2800 Hz—were the lowest scoring conditions for all subjects under all test conditions. Scores for the fourth condition, BP 1400–2800 Hz, were higher than these except for T.T., who did better on the BP 700–1400-Hz condition.

These results show that if the presentation level is sufficiently high, the upper and lower extremes of the available speech bandwidth were not essential for good speech intelligibility of the test materials used in this study for any of the listeners tested. However, removing spectral components below 700 Hz had a lesser effect on reducing intelligibility at higher presentation levels than removing components above 2800 Hz. It also seems apparent from these results that the 1400–2800-Hz band contributes the major portion of the intelligibility of the CV syllables used in the present research. The results of the experiments with normal-hearing listeners tested with simulated hearing loss show that for these two filter conditions, they performed more like the listeners with flat sensorineural hearing loss than like those listeners whose hearing loss was predominantly in the higher frequencies.

Intelligibility at High Presentation Levels

The performance-intensity functions for all subjects exhibited roll-over at high presentation levels, defined as a reduction in score of more than three percentage points below the maximum score at a presentation level higher than that at which the observed maximum occurred. Table 4 indicates the instances for which roll-over was observed, based on this definition. A third indication, designated by "-", means that scores at presentation levels higher than that at which the maximum intelligibility occurred were lower than the maximum, but by less than three percentage points. In considering these data it is important to note that while impaired listeners were tested throughout their useful dynamic range, the normal-hearing listeners were generally not tested at comparably high presentation levels. Further, it is possible that even though roll-over was not seen in all cases, it is likely that it eventually would have occurred as level increased.

For the normal-hearing subjects listening in quiet, roll-over did not occur consistently for each subject and each condition. When the normal-hearing subjects listened with the masking noise to simulate hearing loss, roll-over occurred in the unmasked cases, the narrow band-pass filters and low-pass 700-Hz filtered speech. However, as level increased to overcome masking, with only one exception, roll-over did not occur in the wider band condition as it did for the impaired subjects.

Roll-over occurred more frequently for the impaired listeners than for the normal-hearing listeners. For subject F.G. no obvious pattern appeared in the conditions that exhibited roll-over, although roll-over was observed for more conditions in the right or better ear. This is contrary to what one would intuitively expect, that is, that saturation or roll-over would occur in the ear with the more severe hearing loss. For subject G.M. roll-over took place for every condition tested in the left ear and for 5 of the 10 conditions in the right. However, in the right ear roll-over did not occur in a predictable manner. In G.M.'s left ear roll-over for the LP 1400-Hz condition occurred at a relatively low presentation, as it did for F.G. In the observed performance for J.G.'s impaired ear roll-over was observed for two conditions, HP 2800 Hz and BP 700–2800 Hz. In particular, the roll-over

TABLE 4. Roll-over conditions for subjects.

	Normal-hearing			*Hearing-impaired*						*Simulated loss*		
	J.G.	*J.T.*	*R.M.*	*F.C.*		*G.M.*		*J.G.*	*T.T.*	*J.G.*	*J.T.*	*R.M.*
Filter	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*R*	*L*	*R*	*R*
Unfiltered	Y	Y	–	Y	Y	Y	Y	N	N	N	N	Y
LP 700	N	Y	Y	Y	–	–	Y	–	–	N	Y	Y
LP 1400	–	N	N	Y	Y	Y	Y	N	N	N	Y	Y
LP 2800	N	N	N	N	Y	–	Y	N	Y	N	N	N
HP 700	N	–	N	Y	–	Y	Y	N	–	N	N	N
HP 1400	–	N	N	Y	Y	–	Y	–	Y	Y	N	N
HP 2800	N	N	Y	Y	Y	N	Y	Y	Y	N	Y	–
BP 7–14	Y	N	Y	Y	–	Y	Y	N	N	Y	Y	Y
BP 14–28	–	–	–	–	–	Y	Y	–	–	Y	Y	Y
BP 7–28	Y	N	–	Y	Y	N	Y	Y	N	N	N	N

Note. LP = low-pass; HP = high-pass; BP = band-pass.

was especially great for BP 700–2800 Hz. For several other conditions the drop in score was very slight. For subject T.T. roll-over occurred for three conditions. Since she was the only listener who regularly wore a hearing aid, it is possible that she may have a greater tolerance for speech than the other listeners.

Several hypotheses exist for the presence of roll-over. One is the masking of high-frequency energy in speech caused by the upward spread of masking from relatively intense low-frequency energy that may reduce intelligibility as intensity increases (French & Steinberg, 1947; Kryter, 1962a). In this study, spread of masking effects may have caused roll-over at levels below discomfort thresholds for the low-pass conditions. For the hearing-impaired listeners with high-frequency sensorineural loss, low-level consonant energy is crucial to intelligibility. The spread of masking effects will likely have greater influence on intelligibility because the masking extends into regions where listeners have impaired hearing mechanisms.

Another reason for roll-over is the possible increase in internal distortion, especially in impaired auditory systems, caused by high-intensity speech near discomfort thresholds. Because speech consists of a distribution of sound levels, brief peak speech levels may have been associated with such internal distortion, but not for a long enough time period that a subject experienced discomfort. It is important to realize the difficulty in accurately specifying the speech levels. Since root-mean-square speech levels were referred to a sine wave level in calibration, it is possible that for some of the speech signals used in this study, peak values exceeded discomfort thresholds.

Roll-over occurred more often for the impaired subjects than for the normal-hearing subjects. However, as with the normal-hearing subjects, there was no discernable pattern to its occurrence. The results obtained for subject F.G. demonstrate the apparent effects of the spread of masking and increased distortion. For F.G. roll-over occurred at a lower level for unfiltered speech than for HP 700 Hz when the low-frequency energy is removed by filtering. His performance at high levels for HP 700 Hz improved relative to unfiltered speech, indicating that he was able to make use of his residual high-frequency hearing in the absence of intense low-frequency energy.

Roll-over is seen in the data of French and Steinberg (1947) for nearly every condition of filtered speech plotted by them. They attributed this to "a fatigue effect which may be considered as self-masking" (p. 109). At the highest level reported by them, intelligibility performance was lower for all conditions. In the present study, measurements of the normal-hearing listeners tested at a comparable level showed no consistent pattern of roll-over. The reader must be cautioned that these data were obtained from only three listeners whose age varied greatly. In most studies with normal-hearing listeners, the subjects are usually greater in number and more homogeneous as a defined group.

None of Pollack's (1948) smoothed "gain functions" exhibited roll-over or reduced performance. Only one data point indicated by Pollack appeared to show a reduction in performance at the highest tested level. This is surprising since his highest presentation levels seem to be more intense than the highest levels in this study. For the subjects in the present study, roll-over in the performance-intensity functions occurred for some conditions at levels substantially below Pollack's.

Other Studies of Filtered Speech Perception

Comparing the results of this study to other studies is difficult because few investigators tested subjects over a wide range of intensities, and even fewer evaluated both normal-hearing and hearing-impaired listeners under identical training and test conditions. Another problem is that, with one exception, speech materials used by the investigators differed considerably from those of the present study. Furthermore, the characteristics of the digital filters used here were not likely to be similar to any of the analog filters used by other investigators. Of those investigations that tested a range of intensities, the reports of French and Steinberg (1947) and Pollack (1948) have relevance to this research.

Since the data reported by French and Steinberg (1947) are often considered to be representative of the performance of normal-hearing subjects listening to filtered speech, it was of interest to make direct comparisons of the results of this study with their results. For this comparison the results for J.T. and the normally hearing left ear of J.G. were combined because they were normal-hearing, relatively young listeners. Figure 5 reproduces the data of French and Steinberg for the filtered conditions closest to those tested in this study, along with the present data to allow direct comparisons of the data. French and Steinberg originally plotted their data relative to the orthotelephonic condition. This was converted to sound pressure level by assuming that 0 dB orthotelephonic gain equals 65 dB SPL.

The unfiltered condition of the present study is actually low-pass filtered at 4500 Hz and should thus be compared to French and Steinberg's (1947) low-pass 4500-Hz condition. For the present unfiltered condition, maximum performance was nearly equal to French and Steinberg's LP 4500 Hz and occurred at approximately the same level. However, as level decreased, scores for the present study dropped somewhat more rapidly than for French and Steinberg. The differences are primarily one of a translation of the curves as a function of intensity since the slopes of the curves are similar. There appears to be a small but systematic difference between the specification of presentation levels used in this study and the levels indicated by French and Steinberg.

Differences in performance similar to the wider band conditions are seen for the LP 2800-Hz and for French and Steinberg's (1947) LP 2850-Hz conditions. The scores reported by French and Stein-

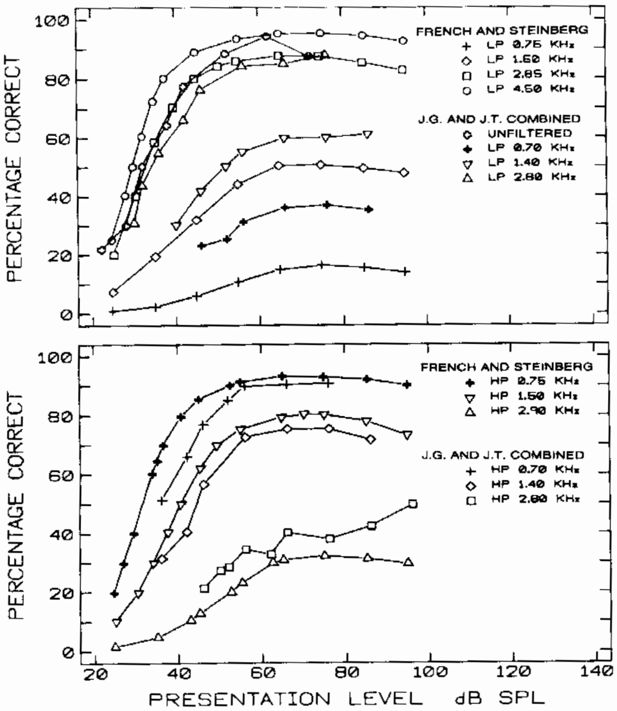

Figure 5. Averaged performance-intensity functions for normal listeners J.G. and J.T. compared to results of French and Steinberg (1947).

berg's are generally higher than those obtained in this study. Again, there appears to be a difference in the specification of presentation levels since the slopes of the curves are similar. However, for the low-pass 1400/1500-Hz and low-pass 700/750-Hz conditions, scores achieved by the listeners in this study were significantly higher than French and Steinberg's, even though the cut-off frequencies were slightly lower and the filter skirts significantly steeper.

Some differences are also evident in the high-pass filter conditions. French and Steinberg's (1947) scores are higher for the HP 700/750-Hz and HP 1400/1500-Hz conditions. As for the low-pass cases, there appears to be a systematic level difference since the slopes of the P-I functions are similar. For HP 2800 Hz, however, scores in the present study were higher and increased above 80 dB SPL. This is most likely due to differences in filter characteristics.

Differences between the results of this study and of French and Steinberg's (1947) may be attributable to several factors. These include differences in procedures, speech levels, speech materials, subject and talker differences, filter characteristics and listening conditions. For the French and Steinberg report the subjects heard the speech signals read by natural voices over a simulated telephone link. They responded verbally or wrote the responses from a virtually unlimited response set. In the present study, the subjects' responses were restricted to the 72 possible choices of the utterance. This may account for the relatively higher performance scores achieved by the listeners in this study.

Another source of variation resulted from the specification and control of speech levels. In this study the unfiltered CVs were normalized to equal vowel RMS levels. It is not likely that any normalization was used by French and Steinberg (1947) other than verbal instructions to the talkers to attempt to maintain a uniform speaking level. In addition, because the subjects in this study were trained extensively on each condition, as reported by Milner (1973), performance levels would be expected to be higher than when subjects were given less rigorous training, unless testing continued over a long period of time with the same subjects.

Differences in the test materials partially account for differences in the data of this study as compared to that of French and Steinberg (1947). The set of CVs employed for this study used only three vowel environments—/a/, /i/, and /u/. The listeners also had to choose one consonant from 24 to successfully identify the syllable correctly. French and Steinberg's subjects listened, however, to CVCs in an open-response format, and therefore, had much lower a priori probabilities of getting equal or higher scores for difficult conditions. In addition, scores for final consonants are usually lower than for initial consonants under difficult listening conditions, further decreasing the intelligibility score.

Differing filter characteristics also contributed to differences between the present study and French and Steinberg (1947) in the results for the high-pass filter conditions. The digital filters used in this study had extremely sharp slopes. Since digital filters attenuate linearly as a function of frequency rather than logarithmically, the slope characteristics are different as well. French and Steinberg used standard analog filters whose slopes were unspecified, but not likely as steep as the digital filters. In addition, the speech signals passed through a 4500-Hz anti-aliasing filter with a cut-off slope of approximately 120 dB/octave. As Castle (1963) and Palva (1965) observed, the slopes of the filter skirts are a critical variable in the perception of filtered speech. Another characteristic of the digital filters is the absence of phase distortion. This may also have had a slight influence on the results of this study, tending to improve scores relative to similar analog filters where phase distortion may be significant.

Pollack's (1948) study may be compared to the results obtained here for the normal-hearing listeners tested with masking noise to simulate hearing loss. Using monosyllabic phonetically balanced words, Pollack tested normal-hearing listeners with several high-pass and low-pass filtered speech conditions in the presence of a constant-intensity white-noise signal. His noise level was 10–15 dB more intense than the maskers used in the present study. At the highest levels tested, however, maximum intelligibility scores were comparable for comparable conditions. Pollack found that maximum performance for masked high-pass filtered speech with a cut-off of 350 Hz was higher than that for the unfiltered condition, similar to the results observed for the subjects in this study listening to the HP 700-Hz condition. At Pollack's next highest high-pass cut-off—580 Hz—the effect disappeared, consistent with the performance of the normal listeners tested with the noise-simulated hearing loss. Pollack also observed that extreme high-frequency components are important. This was confirmed by the improved performance for the conditions when energy above 2800 Hz is added to the low-pass and band-pass conditions with an upper cut-off frequency of 2800 Hz.

Pollack observed differences in the "gain functions" of the high-pass and low-pass filter conditions, noting that the slopes of the high-pass functions were much steeper than the low-pass functions. A similar effect was seen in this study, where one may observe that the minimum scores for the high-pass conditions occurred at significantly higher presentation levels than for the low-pass conditions and rose sharply to maximum performance levels comparable to the levels in the unmasked cases.

LaBenz (1956) used monosyllabic phonetically balanced words to test 100 adult subjects with a nearly uniform mix of normal-hearing, conductive, mixed, and sensorineural ("perceptive") hearing losses. All testing was done at 30 dB relative to spondee threshold (SRT). Relative to the normal-hearing, conductive, and mixed loss subjects, averaged data obtained from the subjects with sensorineural loss scored lower as cut-off frequencies increased in low-pass conditions. For band-pass conditions consisting of varying widths with cut-off frequencies ranging from 250 Hz to 3000 Hz, highest scores for all listeners were achieved for the bands of 1500–2000 Hz and 2000–3000 Hz. However, performance of the sensorineurally impaired subjects was substantially poorer than for the others. Compared to the results of the present study, LaBenz's sensorineurally impaired subjects performed much more poorly than the impaired subjects used here, except for T.T. and the right ear of G.M.

LaBenz (1956) claimed that the subjects with sensorineural hearing loss performed similarly to normal-hearing listeners hearing low-pass filtered speech. The results of the present study agree with this finding as far as comparisons are possible. Generally, the subjects with high-frequency sloping losses tested for the present research had reduced intelligibility scores for the wide-band condition, compared to the normal listeners. Their maximum scores fell in a range where the results for a low-pass condition might have fallen had such a condition been tested. LaBenz did not test high-pass conditions, so one cannot see if performance improved for high-pass conditions with low cut-off frequencies relative to wide-band speech.

Castle (1963) studied normal-hearing listeners using monosyllabic words read by a single male talker with band-pass filters of varying width and filter skirt slopes. Only three of Castle's many conditions match closely to those studied here. They were band-pass 720–1440 Hz, 1440–2400 Hz, and 720–2400 Hz. The scores achieved by Castle's subjects for these conditions were dependent on the filter skirt configuration. The maximum scores for the comparable conditions in this study were somewhat lower than Castle's, but the relative ordering of performance was the same: BP 700–1400 Hz was the poorest, next highest was BP 1400–2800 Hz, and highest was BP 700–2800 Hz.

Palva (1965) studied the performance of normal-hearing subjects listening to filtered Finnish two-syllable words. For Palva's high-pass conditions, scores were generally lower than those in the present study at comparable filter cut-off frequencies, whereas for low-pass conditions, scores were higher. These differences were due, most likely, to the greater vowel content of the Finnish words that Palva used. The band-pass filter conditions he studied increased in bandwidth steps from extremely narrow to 1½-octave-wide filters. Band center frequencies varied from about 300 to 5000 Hz. The one octave case is directly comparable to the band-pass filters BP 700–1400 and BP 1400–2800 Hz, studied for this research. Palva's results were the reverse of the outcome of this study, with higher scores for the lower octave band. Palva's overall maximum score occurred for the bands 900–1800 Hz and 1080–2160 Hz. Again, this result was most likely due to the nature of the speech materials.

Thomas and Pfannebecker (1974) tested three conditions of high-pass filtered speech using hearing-impaired subjects. The three conditions consisted of Harvard PB-50 words (Egan, 1948) filtered at 1600 Hz, with three slope characteristics of 12, 18, and 24 dB per octave. They theorized that these gradual slopes would alter the ratio of the first and second formants (*F*1 and *F*2) to permit greater audibility of *F*2 and improve intelligibility, especially for listeners with high-frequency hearing loss. Their data support the notion that reducing the amount of low-frequency energy in speech results in improved intelligibility for listeners with primarily high-frequency sensorineural hearing loss.

The study by Wang, Reed, and Bilger (1978) is similar to this study in that both used digitally recorded nonsense syllables without a carrier phrase. However, Wang et al. employed both consonant-vowel (CV) pairs and vowel-consonant (VC) items, but used only one presentation level for the eight normal-hearing listeners. The talker of these materials was unspecified and only one token per syllable was presented, as contrasted to four for this study. Wang et al. presented averaged curves of percentage correct as a function of filter cut-off frequency. As far as one may compare, the data of the studies appear similar except at the extreme ranges of high-pass and low-pass filtering. The scores reported by these authors were higher for these cases. These differences are possibly due to the differences between their analog filters and the digital filters of the present study. The authors made no mention of waveform amplitude normalization, as was done on the speech signals for this study.

CONCLUSIONS

1. For the normal-hearing subjects listening in quiet and noise and the hearing-impaired subjects listening in quiet who participated in this study, speech intelligibility of the materials used remained relatively high whether high-pass filtered at 700 Hz or low-pass filtered at 2800 Hz, when heard with sufficient intensity.

2. For all the listeners in this study, roll-over in performance occurred at high presentation levels that were below reported discomfort thresholds. Identification of when that might occur was not specifically clear from the results of this study.

3. For listeners with sloping sensorineural hearing loss, removing the high-frequency energy from 2800 Hz up to the 4500-Hz cut-off of the material presented here had little effect on reducing intelligibility. A more significant effect resulted from removing the energy below 700 Hz. At sufficiently high presentation levels, intelligibility performance usually equaled or exceeded that achieved for the unfiltered condition incorporating the spectral components below 700 Hz.

4. The performance-intensity functions of the normal-hearing listeners with simulated high-frequency hearing loss were similar to listeners with relatively flat sensorineural hearing loss. Eliminating low-frequency energy for these subjects did not have as dramatic effect on their intelligibility performance as it did on the impaired subjects with mainly high-frequency hearing loss. That is, the simulation of sloping loss in the normal-hearing listeners was more like a flat loss rather than a sloping loss.

ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions of D. K. Bustamante, M. C. W. Coln, S. V. DeGennaro, D. M. Freeman, P. W. Herman, M. A. Picheny, and P. M. Peterson who developed hardware and software used in this research. Support for the research was provided by the National Institutes of Health (Grants R01 NS 12846 and T32 NS 07099).

REFERENCES

BILGER, R. C., & HIRSH, I. J. (1956). Masking of tones by bands of noise. *Journal of the Acoustical Society of America, 28*, 623–630.

CASTLE, W. E. (1963). Effects of selective narrow-band filtering on the perception by normal listeners of Harvard PB-50 Word Lists. *Journal of Speech and Hearing Association of Virginia, 5*, 12–21.

EGAN, J. P. (1948). Articulation testing methods. *Laryngoscope, 58*, 955–981.

FRANKLIN, B. (1969). The effect on consonant discrimination of combining a low-frequency passband in one ear with a high-frequency passband in the other ear. *Journal of Auditory Research, 9*, 365–378.

FRANKLIN, B. (1975). The effects of combining low- and high-frequency passbands on consonant recognition in the hearing impaired. *Journal of Speech and Hearing Research, 18*, 719–727.

FRENCH, N. R., & STEINBERG, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19*, 90–119.

HIRSH, I., REYNOLDS, E. G., & JOSEPH, M. (1954). Intelligibility of different speech materials. *Journal of the Acoustical Society of America, 26*, 530–538.

KRYTER, K. D. (1962a). Methods for the calculation and use of the Articulation Index. *Journal of the Acoustical Society of America, 34*, 1689–1697.

KRYTER, K. D. (1962b). Validation of the Articulation Index. *Journal of the Acoustical Society of America, 34*, 1698–1702.

LABENZ, P. J. (1956). *Potentialities of auditory perception for various levels of hearing loss* (Volta Bureau Reprint 683). Washington, DC: A. G. Bell.

LIPPMANN, R. (1981). MX41/AR Earphone cushions versus a new circumaural mounting. *Journal of the Acoustical Society of America, 69*, 589–592.

McCLELLAN, J. H., PARKS, T. W., & RABINER, L. R. (1973). A computer program for designing optimum FIR linear phase digital filters. *IEEE Transactions on Audio and Electroacoustics, AU-21*, 506-526.

MILLER, J. D., ENGEBRETSON, A. M., GARFIELD, S. A., & SCOTT, B. L. (1975). New approach to speech-reception testing. *Journal of the Acoustical Society of America, 57*(Suppl. 1), S48.

MILNER, P. (1973). Advantages of experienced listeners in intelligiblity testing. *IEEE Transactions on Audio and Electroacoustics, AU-21*, 161–165.

PALVA, A. C. (1965). Filtered speech audiometry, I. Basic studies with Finnish speech toward the creation of a method for the diagnosis of central auditory disorders. *Acta Otolaryngologica*, (Suppl. 210).

POLLACK, I. (1948). Effects of high-pass filtering on the intelligibility of speech in noise. *Journal of the Acoustical Society of America, 20*, 259–266.

ROSENTHAL, R. D., LANG, J. K., & LEVITT H. (1975). Speech reception with low frequency speech energy. *Journal of the Acoustical Society of America, 57,* 949–955.

THOMAS, I. B., & PFANNEBECKER, G. B. (1974). Effects of spectral weighting of speech in hearing-impaired subjects. *Journal of the Audio Engineering Society, 22,* 690–694.

VILLCHUR, E. (1970). Audiometer-earphone mounting to improve intersubject and cushion-fit reliability. *Journal of the Acoustical Society of America, 48,* 1387–1396.

WANG, M. D., REED, C. M., & BILGER, R. C. (1978). A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusions. *Journal of Speech and Hearing Research, 21,* 5–36.

Chapter 7

CLINICAL APPLICATIONS OF COMPUTER-AIDED TECHNIQUES FOR SPEECH RECOGNITION TESTING

CANDACE A. KAMM*

AT&T Information Systems Laboratory
Lincroft, NJ

Currently, computer-aided techniques for evaluating speech recognition of hearing-impaired individuals are not implemented widely in audiology clinics. Laboratory experience indicates that such techniques may increase the efficiency and accuracy of speech recognition testing. The focus of this presentation is toward clinical applications of computer-assisted methods for testing speech recognition. This presentation reviews digital speech processing and compares a sample application (an adaptive procedure) using computer methods and digitized speech materials to a manual method using analog recordings. Finally, the requirements for implementation of computer-assisted test procedures are defined.

Digitized Speech Materials

In this presentation, the term *digitized speech materials* is limited to the digital representation of a real-speech waveform. That is, synthetic speech stimuli are not considered, as there are currently no speech recognition tests using such stimuli. Furthermore, the stored data are assumed to be a digital representation of the analog waveform rather than a more reduced representation, such as linear predictive coding (Rabiner & Schafer, 1978).

Figure 1 provides a brief review of the process of converting a continuous signal into a digital signal. The electrical analog signal is low-pass filtered to limit the bandwidth of the signal to frequencies less than half the sampling rate. This filtering is required because only those frequencies less than or equal to one half the sampling rate can be reconstructed uniquely from a digital representation. Thus, if a particular bandwidth signal is desired, the sampling rate and filtering requirements are essentially determined. After filtering, the signal is passed through an analog-to-digital converter (A/D), which samples the analog signal every *t* seconds and quantizes the amplitude to *n* bits. The number of bits in the A/D converter defines the number of discrete values that are available for estimating the instantaneous voltages of the sampled waveform. The number of bits determines the resolution of the digitizing process and so is central to the issue of signal-to-noise *(S/N)* ratio of the digitized speech. To optimize *S/N* ratio, the signal level used during the digitization process should approach the voltage range of the A/D converter. For each sample, the A/D conversion process results in a digital word that provides the closest match to the level of the analog signal at each sampling point. The digital representation of the signal is stored as a sequence of binary numbers in computer memory and later stored on a mass storage device. To recover the analog signal, the above process is reversed.

There are several advantages and limitations to using digitized speech stimuli for speech recognition testing. Among the advantages of digitized speech materials are the following:

1. The stored sample will not deteriorate in quality over time, either due to storage or playback.
2. The signals can be manipulated using computer algorithms. For example, digitized signals can be time compressed, filtered digitally, edited, concatenated, and so on.
3. Perhaps most useful to clinical applications, the speech items can be accessed in random order.
4. Physical characteristics of the signal can be specified precisely.

There are some potential limitations of digitized materials, as far as clinical applications are concerned. These include the following:

1. There may be storage limitations for large open-set tests [e.g., the revised SPIN sentences (Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984) consist of eight 50-item lists, each item approximately 1.5 s long]. The storage capability of the computer disk would need to hold 15 million samples at a rate of 25 kHz to store the entire test.
2. Another more subtle problem involves the issue of maintaining form equivalence for tests with time-locked competition (e.g., the SPIN). The digitizing process just described was directed toward representing only the primary signal. To provide the required time locking of primary and masking signals, a second A/D converter and some intricate timing algorithms would be necessary for digitization, and concomitant dual D/A conversion would be required for playback.

It is clear that the digitization is to some extent a reduction in information relative to the continuous analog signal. How-

[*] *Currently at Bell Communications Research, Murray Hill, NJ.

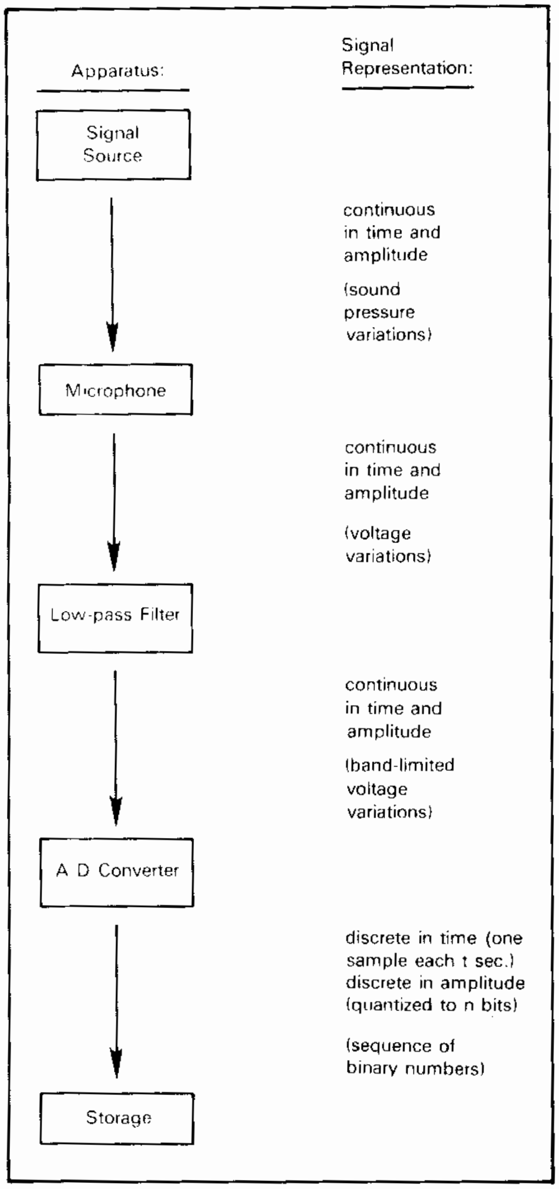
Apparatus:
Signal Representation:
Signal Source
continuous in time and amplitude
(sound pressure variations)
Microphone
continuous in time and amplitude
(voltage variations)
Low-pass Filter
continuous in time and amplitude
(band-limited voltage variations)
A D Converter
discrete in time (one sample each t sec.)
discrete in amplitude (quantized to n bits)
(sequence of binary numbers)
Storage

FIGURE 1. Analog-to-digital (A/D) conversion.

ever, with the availability of 16-bit D/A converters and the use of sampling rates on the order of 25 kHz, excellent speech quality with bandwidth 10 kHz is practical.

Sample Clinical Application: Comparison of Analog and Digital Methods

Once the speech items are stored in digital form, they can be incorporated in procedures that would increase efficiency and accuracy of clinical testing. One example of a task well-suited for computer methods is the adaptive procedure discussed in the chapter by Dr. Levitt. There are several steps involved in the analog form of such a procedure designed to converge on the sound pressure level at which a criterion level of speech recognition performance (e.g., 50%) is obtained. For the purposes of this example, assume that a closed-response test like the Nonsense Syllable Test (NST) is used (Resnick, Dubno, Hoffnung, & Levitt, 1975). For the manual analog version of this procedure, the clinician first turns on a tape recorder and sets appropriate attenuation to start the test. It is essential that the clinician have in mind the adaptive transform (Levitt, 1971) required, including the criteria for changing attenuation and for terminating the procedure. The audiologist listens to the stimulus, listens for patient's response, records the response, judges whether the response is correct, counts reversals of intensity change of the procedure to determine when the procedure is completed, and if completed, turns off the tape recorder and computes the average level of the midpoints of the adaptive runs. If the test is not completed, the clinician must decide whether the adaptive rule requires an attentuation change and, if so, must change the attenuator prior to the presentation of the next word. Keep in mind that the pause between items on most commercially available test tapes is approximately 5 s. With practice, a simple up-down adaptive procedure to determine the sound pressure level required for 50% performance is a feasible clinical tool, even in this analog form. However, the procedure can be facilitated using computer methods.

Certain tasks are required of the audiologist during the test, and certain activities are computer controlled. First, the audiologist sets up the adaptive procedure parameters, such as starting level, step size, stopping criterion, and adaptive transform (i.e., what criterion performance the rule should converge on). The computer program then sets the attenuation. The speech items are accessed in a pseudo-random fashion without replacement. To economize on storage, carrier phrases common to all test items may be stored separately from the items and combined prior to playback. In this example, single tokens of the initial NST carrier "you will mark" and the final NST carrier "please" could be stored on disk, along with the 91 test syllables in isolation. The carriers and randomly selected item are concatenated into a natural sounding utterance and played to the patient, who indicates the response for each item on a response box. The program checks for the accuracy of the response by comparing the response code with the stored stimulus code, and checks the rule for termination of the procedure. If the termination criterion has been realized, statistics (e.g., average levels of midpoints of the test runs) are computed and printed in hard copy form for the audiologist. If the procedure is not complete, the computer program checks the adaptive rule, determines whether an attenuation change is necessary, and the process continues. Meanwhile, the clinician monitors responses passively and checks the progress of the procedure to identify any difficulties.

Obviously, this example of an adaptive procedure using a closed-response-set test is the most advantageous case in favor of computer methods, unburdening the clinician of monitoring intensity-changing and stopping rules, as well as handling response collection. Other uses include (a) more typical playback capabilities used in routine recognition testing (i.e., single intensity, random order), (b) testing of speech recognition in noise using adaptive procedures where the level of noise is altered to determine the *S/N* ratio required for a criterion performance level, and (c) estimating the signal level at which maximum recognition would be achieved (again using adaptive procedures). These methods can also be used with open-response-set tests by requiring the audiologist to indicate the accuracy of the subject's responses to each item.

Implementation of Computer Methods in Clinical Settings

Implementing a system for computer-assisted testing of speech recognition performance using digitized speech materials involves the purchase of a microprocessor, mass storage device, and other peripherals; a method of obtaining the stimuli (either by digitizing items from an analog source or obtaining item files in digital form from a vendor); programming support and maintenance; and training for non-computer-oriented clinicians.

At the present time, no *clinic* has implemented a system for presenting digitized speech. However, Drs. Gerald Popelka and A. Maynard Engebretson (1983) are currently developing a computer for speech audiometry at the Central Institute for the Deaf that will deliver digitized speech stimuli. According to Dr. Popelka, the microprocessor-controlled HP-85 lists for approximately $2,500 and comes equipped with a printer for hard copy and a terminal and keyboard. The mass storage devices are two Winchester disk drives (cost, approximately $2,500) with 38 Mbytes of storage, enough to contain approximately 10 min of speech samples at 50 kHz. He plans to store a number of speech tests, including the California Consonant Test (Owens & Schubert, 1977), the NST, and the Pascoe High-Frequency Word List (Pascoe, 1975). Dr. Popelka has the advantage of being in an environment with an active research staff, and thus has resources available for building many of the necessary peripheral devices, including the subject response apparatus, A/D converter, and concomitant interfaces. He is also adept at programming, and so has the flexibility to adapt the system to the needs of his clinicians.

It is doubtful whether most clinics have the capability to implement such a system. Digitized speech will most likely be delivered to the audiology clinic through the development of speech audiometers with preprogrammed speech sets and test procedures. Audiometers that are microprocessor-based are currently available from several vendors, although speech recognition testing is not one of the automated modules offered. If quality control and standardization among clinics are desirable, some basic guidelines should be suggested before a myriad of digitized speech sets becomes available. For example, the number of bits in the D/A converter, filter characteristics, and bandwidth requirements are among the variables which would benefit from common agreement.

In addition, another issue should be considered. Studies using analog materials suggest that as redundancy is reduced by eliminating portions of the signal, the performance of hearing-impaired listeners is reduced disproportionately as compared to normal-hearing listeners. Digitization is essentially a data reduction technique. Whether digitization yields different effects on speech recognition of hearing-impaired listeners as compared to normal-hearing listeners is uncertain. Certainly, with a 10-kHz bandwidth and the resolution of a 16-bit D/A converter, performance of listeners with moderate sensorineural hearing loss is nearly identical to their performance using analog materials, at least when listening in quiet. It might be insightful to systematically consider the effects of changing such variables as sampling rate and number of bits on the relative recognition performance of normal-hearing and hearing-impaired listeners before specifying guidelines for digitization of speech recognition materials.

Summary

In summary, it is likely that the use of computer methods for test procedures accessing speech stimuli stored in digital form would increase efficiency and accuracy of many speech recognition tests. However, in part because of cost and also because of lack of experience with this application of computer technology, such procedures are not now in widespread use clinically. There can be little doubt that they will be available soon. Computerized methods will facilitate the use of more standard materials and test procedures among clinics than currently occur using analog signals such as tape recordings and monitored live voice. However, the increased efficiency and accuracy that can be achieved using computer-assisted test procedures will not be of benefit in the clinic until audiologists resolve the more critical issues concerning test validity, with respect to the hearing-impaired individual's communicative ability in everyday listening situations and test reliability of currently available speech recognition materials.

REFERENCES

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of Speech Perception in Noise. *Journal of Speech and Hearing Research, 27,* 32–48.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467–477.

Owens, E., & Schubert, E. D. (1977). Development of the California Consonant Test. *Journal of Speech and Hearing Research, 20,* 463–474.

Pascoe, D. P. (1975). Frequency responses of hearing aids and their effects on the speech perception of hearing-impaired subjects. *Annals of Otology, Rhinology, and Laryngology, 84*(Suppl. 23).

Popelka, G. R., & Engebretson, A. M. (1983). A computer-based system for hearing aid assessment. *Hearing Instruments, 34,* 6-9, 44.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals.* Englewood Cliffs, NJ: Prentice-Hall.

Resnick, S. B., Dubno, J. R., Hoffnung, S., & Levitt, H. (1975). Phoneme errors on a Nonsense Syllable Test. *Journal of the Acoustical Society of America, 58*(Suppl. 1), S114.

Chapter 8

IMPLICATIONS OF THE AUDITORY-PERCEPTUAL THEORY OF PHONETIC PERCEPTION FOR SPEECH RECOGNITION BY THE HEARING IMPAIRED

JAMES D. MILLER

Central Institute for the Deaf
St. Louis, MO

The auditory-perceptual theory of phonetic perception describes, as yet incompletely, a series of transformations of the speech waveform that lead to a neural representation of the phonetic structure of an utterance as a series of phones (perceived speech sounds). This paper presents the basic concepts of the theory.

THE AUDITORY-PERCEPTUAL THEORY

Phonetically Relevant Auditory-Perceptual Space

It is assumed that the phonetically relevant aspects of speech can be revealed by consideration of the power spectra of brief, 10–20-ms segments of the speech waveform. The work of Miller, Engebretson, and Vemula (1980a, 1980b, 1982) implies that when these spectra are expressed in log-power and log-frequency dimensions, or similar loudness-like and pitch-like dimensions, then the absolute position of these spectra along either dimension is relatively unimportant. That is, their sensory effect is controlled by spectral shape rather than spectral position and, within limits, simple transpositions along either dimension will not alter the phonetic information being carried by the signal.

It is further assumed that each "spectral shape," so expressed, can be represented as a single point in a phonetically relevant auditory-perceptual space of only a few dimensions. The concept of such a space has been previously suggested by authors such as Peterson (1952), Shepard (1972), and Pols (1977). One possible prototype for a phonetically relevant auditory-perceptual space (Miller, 1982a, 1982b, 1982c, 1982d) has the dimensions $x = \log(F3/F2)$, $y = \log(F1/F0')$, and $z = \log(F2/F1)$. Here $F1$, $F2$, and $F3$ are taken as the frequency locations of the first three significant spectral prominences of a brief 10–20-ms segment of speech. For periodic speech, $F0' = aF0$, where $F0$ is the fundamental frequency of the voice and a is an arbitrary constant which, on the average, has a value about 1.5 times greater for males than for females or children. For transients or aperiodic speech, $F0'$ is taken as equal to $F1$. This space can be displayed as a physical three-dimensional model (Heidbreder & Miller, 1982) and exhibits many interesting properties, which have been described in another paper (Miller, 1982e).

Sensory Pointer, the Sensory Path, and the Auditory State

Again, it is assumed that the spectral shape of the incoming speech waveform can be represented at any moment as having a position in the phonetically relevant auditory-perceptual space. This position is thought of as being indicated by a very small object, the spectral or *sensory pointer*. As the spectral shape changes in time, this tiny object, the sensory pointer, describes a spectral or *sensory path* through the phonetically relevant auditory-perceptual space. To more fully characterize the incoming speech waveform, another description of the spectral or sensory pointer is needed. This is its *auditory state*. Examples of candidates for auditory states are the auditory-acoustic correlates of nasality, voicing, frication, aspiration, and whisper, or the so-called "source characteristics" of speech. The auditory state of the sensory pointer is represented by its appearance (patterning or color). Thus far in the development of the theory, one imagines the auditory system to be performing a short-term analysis on the speech waveform that allows it to represent, every few milliseconds, the spectral shape and the auditory state of the incoming speech. This sensory processing serves as an input to a higher level perceptual system. The perceptual system integrates sensory information over time, identifies auditory-perceptual events or "sounds," and converts the sensory input into a string of neural symbols or category codes corresponding to the phones of a language.

Perceptual Pointer, the Perceptual Path, and Sensory-Perceptual Dynamics

In the auditory-perceptual theory, the integration of sensory information over time is accomplished in the following way. It is proposed that, in addition to the sensory pointer, there is a *perceptual pointer* that represents the perceptual response at any moment. The perceptual pointer moves through the same auditory-perceptual space as does the spectral pointer and, thereby, traces a *perceptual path*. The perceptual pointer is influenced by a variety of forces that do not influence the sensory pointer. The perceptual pointer is attracted to the sensory pointer by a force that probably increases with their separation as would be the case if they were attached by a spring. The response of the perceptual pointer is made sluggish by assuming that it moves through a *resistive medium* and is also influenced by at least two other forces. These are an assumed *inertial property* of the perceptual pointer itself and a tendency of the perceptual pointer to return to the *null point* of the auditory-perceptual space. The null point represents either silence or a spectrum without significant prominences of the kind indicated later in this paper. The parameters of the theory are to be selected such that the perceptual response exhibits a small amount of overshoot and will reach reasonable vowel locations, when starting from silence, in about 50 ms. These proposed dynamic relations are an attempt to give quantitative expression to the notion that the perceptual system uses data from the sensory input to estimate the most likely course of rapidly changing environmental events.

The idea of a sluggish perceptual system trying to track or follow the trajectories of sensory inputs is familiar as in studies of apparent motion. In the case of speech, such a notion is frequently implied or alluded to and it is explicitly stated in the work of Joos (1948). As presented here, however, the sensory-perceptual dynamics are such that rapid spectral changes can have large effects on the perceptual response because these can result in "a large stretch of the spring" that, in turn, can exert a large force on the perceptual pointer. Since a small amount of overshoot is assumed, the sensory response need only "move rapidly" in the right direction in order to "induce" an appropriate perceptual response. This is consistent with the emphasis Stevens and his coauthors have placed on the importance of rapid spectral changes in the perception of consonants (see especially pp. 22–25 of Stevens & Blumstein, 1981).

Auditory State of the Perceptual Pointer and State Switching

As discussed so far, the location of the perceptual pointer in the auditory-perceptual space gives the "perceptual shape" which is, of course, to be differentiated from the spectral or sensory shape. Also, the perceptual pointer, like the sensory pointer, must have at any moment an auditory state. It is assumed that the auditory state of the perceptual pointer matches that of the sensory pointer but that a certain amount of time is required for *state switching*. For example, if both the sensory pointer and the perceptual pointer are in the frication state and the sensory pointer suddenly switches to the voiced, nonnasal state, it is assumed that several milliseconds are required before the perceptual pointer can switch to the new state. [Previously, it was proposed (Miller, 1982b) that there were four auditory perceptual spaces, four tetrahedrons, so joined to form a pyramid and that the source characteristics of periodic versus aperiodic and nasal versus oral defined the four tetrahedrons. Even though current favor is for the concept of the auditory state for handling the source characteristics, the matter is far from settled.]

Auditory-Perceptual Event

So far, the idea of two small indicators—the sensory pointer and the perceptual pointer—flying around in a common auditory-perceptual space and changing their appearances in some way, perhaps color, to indicate their auditory states has been presented. Now the hypothetical mechanisms whereby the perceptual response becomes one of discrete events and categories is introduced.

Loosely speaking, an *auditory-perceptual event* takes place when a sound is heard. In the current theory, an auditory-perceptual event is said to occur when the behavior of the perceptual pointer meets certain criteria. Three candidate criteria are that such an event occurs when the perceptual pointer (a) undergoes a period of low velocity, (b) undergoes sharp deceleration, or (c) traverses a path of high curvature. It is likely that future research will indicate which of these need to be linked with *and/or* statements, and it is also true that all three may need added constraints, such as a low velocity having to be maintained for *m*-ms or a path of a certain locus and curvature having to be traversed within certain time limits. In any case, this discussion highlights the need to discover those circumstances that lead to an auditory-perceptual event, a discrete sound, rather than a continuously changing, unsegmented flow of auditory experience.

Tick Mark, Clouds of Ticks, Target Zones, Neural Symbols, and Category Codes

The concepts that are intended to aid in understanding the categorical nature of phonetic perception, cross-language differences, and developmental factors in phonetic perception, as well as the selective adaptation experiments, are now introduced.

Each auditory-perceptual event is said to leave a *trace* or *tick mark* that fades in time. When a *cloud of ticks*—that is, a region with a high density of ticks surrounded by a region of lower density—is formed, as would be the case when a stimulus is frequently repeated, it is postulated that the nervous system automatically places an envelope around the cloud of tick marks and creates a *target zone*. Usually, such target zones are temporary and dissolve with time. Other target zones, such as those for the phones of one's native language and dialect, are formed during infancy and childhood under certain circumstances, as yet unspecified. They are nearly permanent and difficult to modify. This may happen in a manner somewhat similar to the sensory imprinting suggested by Marler and Peters (1981).

Target zones are capable of issuing distinct *neural symbols* or *category codes*. An auditory-perceptual event occurring anywhere within a target zone causes it, the target zone, to issue a neural symbol that is determined by the zone itself and by the auditory state of the perceptual pointer at the time of the event. For example, a particular zone may be capable of issuing symbols representing the vowels voiced-/i/, nasal-/i/, and whispered-/i/. Another zone might be capable of issuing symbols corresponding to /b/ and /m/, and so on. Temporary target zones can be formed anywhere in the auditory-perceptual space by presenting a stimulus frequently. These temporary zones may be smaller than and subdivide an existing permanent zone, as seems to be the case in the selective adaptation experiment.

Levels of Processing—The Categorical Level and Several Other Auditory-Perceptual Levels

Liberman and his coauthors (Liberman, 1982, and references cited therein) have repeatedly emphasized that speech-like stimuli may be processed by human listeners in different ways or modes. They refer to a speech mode as opposed to an auditory mode and they rally conclusive evidence in support of this kind of distinction. Below it is suggested that auditory-perceptual information may be processed (a) at the categorical level, that is, in terms of neural symbols or category codes; (b) at the level of the auditory-perceptual event; or (c) at levels of the perceptual and sensory paths. The categorical level is intended to include Liberman's speech mode as well as other nonspeech material that can be similarly processed. The other levels discussed below appear to be subdivisions of his auditory mode.

In the auditory-perceptual theory it is posited that the information from the auditory-perceptual space can be forwarded in various forms to other perceptual-cognitive structures for additional processing. For example, the information can be forwarded as neural symbols issued by target zones which are categorical in nature and place few demands on the listener's perceptual-memorial resources. If more detail is needed, the coordinates of the auditory-perceptual event can be used for additional processing. At an even more detailed level, a highly trained and focused listener may utilize whole segments of the perceptual path, which place considerable demands on his/her perceptual-memorial resources. A fourth possibility is that, under special circumstances of training and attention, a listener may be able to utilize information from the sensory pointer directly.

Role of Top-Down Processing in Phonetic Perception

The theory as just presented does not include the concept of "top-down processing." Indeed, it assumes that when listening to speech that is carefully produced by another native speaker of the same language and dialect no top-down processing is required for accurate phonetic perception. However, the importance of top-down processing in a great many listening situations cannot be denied, and the separation of the perceptual and sensory aspects of phonetic perception leaves ample opportunity for top-down processing to be integrated into the theory.

For example, the listener's expectations could be conceived of as adding forces that attract the perceptual pointer toward particular target zones. In this way, the perceptual pointer is driven not only by the spectral pointer and the other factors previously mentioned in the discussion of sensory-perceptural dynamics, but also by the listener's expectations as they are controlled by context, knowledge of the language, and so on. Another similar form of top-down processing could involve information from other senses resulting in attractive forces on the perceptual pointer. For example, mouth movements such as observed in lipreading could also result in the addition of forces that attract the perceptual-pointer to various target zones and thus influence phonetic perception. Finally, more complicated forms of top-down processing can be imagined. For example, the sizes and shapes of the target zones could be changed depending on the speech characteristics of the talker such as having a foreign accent, deaf speech, and so on. Of course many other kinds of top-down processing can be introduced as the output of the auditory-perceptual space undergoes additional processing such as that required for the identification of words and meanings.

IMPLICATIONS FOR SPEECH RECOGNITION BY THE HEARING IMPAIRED

Failures of the Perceptual Processes

Failures of phonetic perception can now be located in various structures or processes posited in the theory. For example, the locations, shapes, sizes, or functions of the target zones may be faulty as might be the case in certain "higher level" problems such as receptive aphasia, retardation, or prelingual hearing loss. Another possibility, of course, is faulty sensory-perceptual dynamics as sometimes are attributed to some language-impaired children (Tallal, Stark, Kallman, & Mellits, 1981). Perhaps the inertial property of the perceptual pointer is too strong, and the "spring constant" between the sensory and perceptual pointers is too great, resulting in too much overshoot and a generally erratic perceptual response. In contrast, the resistance of the medium of the auditory-perceptual space might be too great resulting in an overly sluggish perceptual response.

Sensory Processes and Their Failures

In the case of uncomplicated, postlingual, sensorineural hearing loss, the sensory-perceptual dynamics and the structures and functions of the target zones are likely to be normal, and it is assumed that the sensory pointer is not moving normally through the auditory-perceptual space.

To better appreciate this possibility, consider in greater detail a scheme for the processing of the spectra of speech that may represent the processing performed by the auditory system. Assume that the auditory system, operating on both place and neural-timing information, performs short-term analyses that (a) give the pitch and pitch strength, (b) identify nasality and other source characteristics, and (c) give the spectral shape. The equivalent computational operations for the determination of spectral shape are as follows. A Fourier spectrum of a brief segment of the speech waveform, such as that given by a 10-ms Kaiser-window, is found and expressed in sensation level (or loudness level) and in log frequency. The spectrum is smoothed by sliding a critical-band-like weighting function along the log frequency or pitch-like axis, and spectral "tilt" is also eliminated by passing the smoothed spectrum through a "high-pass lifter" defined in the log frequency or pitch-like domain. The resulting smoothed spectral envelope is to be rectified to eliminate low-level excursions, including those some fixed number of decibels below the highest spectral peaks as well as those below the threshold of hearing. (These low-level ripples in the processed spectral envelopes are surely irrelevant to phonetic perception.)

Next, the processed spectral envelope would be tested for the presence, location, and strength of an hypothesized "nasal wave" that is suggested by the work of Stevens and Hawkins (1982). Finally, after removal of the effects of nasalization, the resulting spectral envelope is examined for low- and high-frequency cut-offs and significant spectral prominences. These data are then to be used, in a manner not yet fully specified, to define the location of the sensory pointer in a phonetically relevant auditory-perceptual space. Hearing loss, of course, could produce marked aberrations in the resulting spectral shapes. High thresholds, abnormal tilts, abnormal loudness curves, and abnormal frequency resolution (critical bands) will result in incorrect placements of the sensory pointer in the auditory-perceptual space, and the resulting abnormal sensory paths will induce misperceptions. To fully understand this process, it is necessary to know exactly how the abnormal auditory parameters influence phonetic perception.

In the last decade, considerable progress has been made in understanding the speech perception of the hearing impaired through an *acoustical-statistical approach* wherein the distribution of the acoustic energy of speech within the listener's auditory area is examined (Dugal, Braida, & Durlach, 1980; Miller, 1981; Miller, Niemoeller, Pascoe, & Skinner, 1980; Skinner, 1980; Skinner, Karstaedt, & Miller, 1982; Skinner & Miller, 1983; Skinner, Pascoe, Miller, & Popelka, 1982). Although this approach will undoubtedly continue to be important and useful in both the laboratory and the clinic, another approach now appears to offer a new, more detailed and analytic avenue to this area. It is the *phonetic-spectral approach* wherein an attempt is made to learn exactly how the temporal-spectral patterns of speech are processed by the impaired ear and then used to drive the listener's perceptual response to faulty or ambiguous auditory-perceptual events.

REFERENCES

Dugal, R. L., Braida, L. D., & Durlach, N. I. (1980). Implications of previous research for the selection of frequency-gain characteristics. In G. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance and measurement* (pp. 379–403). Baltimore: University Park Press.

Heidbreder, A. F., & Miller, J. D. (1982). Physical models of a phonetically-relevant auditory-perceptual space. In *Periodic Progress Report No. 25*. St. Louis: Central Institute for the Deaf.

Joos, M. (1948). *Acoustical phonetics* (p. 136). Baltimore: Linguistic Society of America.

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist, 37*, 148–167.

Marler, P., & Peters, S. (1981). Birdsong and speech: Evidence for special processing. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives in the study of speech* (pp. 75–112), Hillsdale, NJ: Erlbaum.

Miller, J. D. (1981). Predicting aided speech perception. *Journal of the Acoustical Society of America, 69*(Suppl. 1), S98.

Miller, J. D. (1982a, April). Phonetic perception as an auditory-perceptual process. Lecture delivered at *Perception*, a short course of the Office of Continuing Education of the University of Kansas College of Health Sciences and Hospital.

Miller, J. D. (1982b). Auditory-perceptual approaches to phonetic perception. *Journal of the Acoustical Society of America, 71*(Suppl. 1), S112.

Miller, J. D. (1982c). A phonetically-relevant auditory-perceptual space. In *Periodic Progress Report No. 25*. St. Louis: Central Institute for the Deaf.

Miller, J. D. (1982d). An auditory-perceptual approach to phonetic perception. In *Periodic Progress Report No. 25*. St. Louis: Central Institute for the Deaf.

Miller, J. D. (1982e, November). *A phonetically-relevant auditory-perceptual space*. Paper presented at the 104th Meeting of the Acoustical Society of America, Orlando, FL.

Miller, J. D., Engebretson, A. M., & Vemula, N. R. (1980a). Transposition of vowel sounds. In *Periodic Progress Report No. 23*. St. Louis: Central Institute for the Deaf.

Miller, J. D., Engebretson, A. M., & Vemula, N. R. (1980b). Vowel normalization: Differences between vowels spoken by children, women, and men. *Journal of the Acoustical Society of America, 68*(Suppl. 1), S33.

Miller, J. D., Engebretson, A. M., & Vemula, N. R. (1982). Observations of the acoustic description of vowels as spoken by children, women, and men. *Journal of the Acoustical Society of America, 68*(Suppl. 1), S33.

Miller, J. D., Niemoeller, A. F., Pascoe, D. P., & Skinner, M. W. (1980). Integration of the electroacoustic description of hearing aids with the audiologic description of clients. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 355-377). Baltimore: University Park Press.

Peterson, G. E. (1952). The information bearing elements of speech. *Journal of the Acoustical Society of America, 24*, 629–637.

Pols, L. C. W. (1977). *Spectral analysis and identification of Dutch vowels in monosyllabic words*. Soesterberg, Netherlands: Institute for Perception TNO.

Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David & P. Denes (Eds.), *Human communication: A unified view* (pp. 67-113). New York: McGraw-Hill.

Skinner, M. W. (1980). Speech intelligibility in noise-induced hearing loss: Effects of high-frequency compensation. *Journal of the Acoustical Society of America, 67*, 306–317.

Skinner, M. W., Karstaedt, M. M., & Miller, J. D. (1982). Amplification bandwidth and speech intelligibility for two listeners with sensorineural hearing loss. *Audiology, 21*, 251–268.

Skinner, M. W., & Miller, J. D. (1983). Amplification bandwidth and intelligibility of speech in quiet and noise for listeners with sensorineural hearing loss. *Audiology, 22*, 253–279.

Skinner, M. W., Pascoe, D. P., Miller, J. D., & Popelka, G. R. (1982). Measurements to determine the optimal placement of speech energy within the listener's auditory area: A basis for selecting amplification characteristics. In G. A. Studebaker & F. H. Bess (Eds.), *The Vanderbilt hearing-aid report* (Monographs in Contemporary Audiology, pp. 161–169). Upper Darby, PA: Associated Hearing Instruments.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 1–38). Hillsdale, NJ: Erlbaum.

Stevens, K. N., & Hawkins, S. (1982). Acoustic and perceptual correlates of nasal vowels. *Journal of the Acoustical Society of America, 71*(Suppl. 1), S76.

Tallal, P., Stark, R., Kallman, C., & Mellits, D. (1981). A reexamination of some nonverbal perceptual abilities of language-impaired and normal children as a function of age and sensory modality. *Journal of Speech and Hearing Research, 24*, 351–357.

Chapter 9

DIAGNOSTIC APPLICATIONS OF SPEECH RECOGNITION

DEBORAH HAYES

Baylor College of Medicine
Houston, TX

Diagnostic applications of speech recognition testing include any task employed to identify, evaluate, or rehabilitate the hearing-impaired. Well-defined applications of speech reception testing for diagnosis include identification of presence of a hearing impairment, localization of site of auditory disorder, evaluation of hearing handicap, and evaluation of potential for benefit from auditory rehabilitation. The goal of speech reception in these specific diagnostic applications is to improve treatment of individual hearing-impaired patients.

The origins of diagnostic speech reception measures are closely linked to speech tests developed for communication engineering purposes. Theory, terminology, and even the speech tests themselves developed for engineering systems evaluation were transferred directly to audiology for evaluation of human hearing. At that time, the transfer was appropriate and necessary. The unfortunate consequence of that transfer has been the identification of single-word intelligibility as the de facto standard for diagnostic speech reception measures. The need for a broader spectrum of speech materials is compelling. Essentially, the appropriate speech material for any given diagnostic purpose depends on the level of human communication under investigation. Tests that are appropriate for defining effects of cochlear dysfunction on phoneme identification are inadequate for evaluation of benefit from a cochlear prothesis. Without a doubt, the single most important research need in diagnostic speech reception testing today is the development and validation of a hierarchy of speech materials ranging from purely analytic (i.e., tests of phoneme identification) to largely synthetic (i.e., tests of sentence comprehension). Use of any given material for diagnosis will depend on the nature of the communication behavior under evaluation.

As originally conceived, two speech reception measures described an individual patient's hearing impairment. One measure identified threshold of hearing for speech and provided confirmation of the pure-tone audiogram. The importance of this application has been necessarily diminished by development of sophisticated nonbehavioral estimates of threshold sensitivity. The second measure described the "discrimination loss," or decrease in word intelligibility. This measure served a number of purposes including differentiating conductive from cochlear impairment, estimating degree of hearing handicap, and defining maximum potential for benefit from amplification. Based on current knowledge, the significance of loss in word intelligibility is probably limited to the first application, differentiation of conductive and cochlear impairment.

The next development was use of speech reception measures to evaluate central auditory processes. The main thrust of those efforts was in diagnosis of central nervous system pathology. Two important concepts resulted from this period: first, the recognition of the need for a variety of speech materials to evaluate specific aspects of human auditory function (i.e., identification vs. integration), a need which remains largely unmet; and second, recognition of the role and importance of redundancy of both speech materials and the auditory channel. It is perhaps ironic that the original concept behind speech reception measures in this application—site localization of auditory disorders—is better achieved today by nonbehavioral and even nonauditory measures.

The failure of traditional diagnostic speech reception measures to evaluate social adequacy, or alternatively auditory handicap, are best evidenced by the observation that the single best index of handicap is pure-tone average, or level of threshold sensitivity. Estimates based on word intelligibility do not adequately predict self-reported difficulty in everyday living.

At the present time, then, diagnostic speech reception testing for identification of presence of hearing impairment and localization of site of auditory disorder have limited impact. The most challenging applications of diagnostic speech reception testing for the future lie in three distinct areas: speech reception testing for differentiation of peripheral and central auditory effects in the elderly; speech reception testing in the evaluation of potential for and success of rehabilitation of the profoundly hearing-impaired; and speech reception testing for selection of amplification devices for the mild-to-severely impaired.

The ultimate goal of diagnostic speech reception measures in the first application, differentiation of peripheral versus central effects in the elderly, is improved techniques for rehabilitation of presbycusis. Discrepancy between apparent potential and actual benefit from hearing aid use in the elderly is closely related to presence of significant age-related central effects. By current techniques we are capable of only grossly differentiating peripheral from central effects in this population. The available techniques for evaluation of central auditory effects have met with only limited success in the elderly for three reasons. First, traditional diagnostic speech reception tests were developed to tap a particular level of performance, that is, dichotic listening to evaluate temporal lobe dysfunction. The diffuse and nonspecific nature of central age-related effects displays conflicting patterns of results in many elderly patients. Second, peripheral sensitivity loss accompanying the aging process complicates administration of the test and interpretation of the test result. Finally, and perhaps most importantly, it is often difficult to separate purely auditory effects from more general, nonauditory impairments. For example, test-taking factors such as attention, motivation, memory, and recall are affected by age. These nonauditory effects may produce a pattern of apparent central auditory dysfunction in elderly patients quite apart from actual central auditory effects. Central auditory tests, especially those of complex, "sensitized," or redundancy-restricted materials may be, to a greater or lesser degree, contaminated by these age-related, nonauditory effects. One promising area of exploration appears to be the use of a variety of speech materials, each sensitive to a limited variety of auditory effects.

At Baylor College of Medicine, a group of investigators have compared performance of elderly patients for monosyllabic phonemically balanced (PB) words in quiet to performance for sentences in the presence of ipsilateral speech competition. The rationale underlying this approach is that although no speech recognition task reflects the effects of peripheral impairment exclusive of central function and vice versa, word intelligibility is more sensitive to peripheral than to central effects and sentence identification in competition is more sensitive to central than to peripheral effects. Comparison of performance for the two sets of materials will yield both an estimate of the effects of age-related sensitivity loss on word intelligibility and an estimate of the "peripheral/central ratio" of effects on speech recognition. This approach grew out of the observation that (a) elderly subjects exhibited greater performance deficits for sentences than for words, unlike younger subjects with similar pure-tone audiograms, who exhibit greater performance deficits for words than for sentences; and (b) the nature of this performance deficit (sentences poorer than words) was similar to that exhibited by younger normal-hearing patients with documented central auditory pathology. On the average, this approach provides useful information; however, further refinement and validation of the technique is needed. For example, definition of the contribution of brainstem and temporal lobe components to the specific central age-related effect is potentially useful information. Specific recommendations for auditory rehabilitation may vary depending on the relative contributions of these two central components.

A related research need in this population is definition of the precise effects of central age-related auditory dysfunction on successful use of a hearing aid. Specifically, does successful hearing aid use covary with the relative contributions of peripheral and central effects in the speech recognition problems of the elderly? Is there a specific level of peripheral loss at which amplification is useful, irrespective of integrity of central auditory function? And even more basically, is the peripheral/central ratio of age-related auditory dysfunction a useful predictor of successful hearing aid use?

The second application of diagnostic speech reception measures, evaluation of potential for and benefit from auditory rehabilitation of the profoundly hearing-impaired, developed from the current technology of cochlear prostheses. It is a particularly challenging problem because we have no suitable test materials for evaluation of this subgroup of patients. Investigators quickly found out that open-message set tests of single-word intelligibility were useless in evaluation of subjects with severely restricted auditory systems. The challenge of this problem, then, is to develop measurement techniques of a sufficient range of listening difficulty which are sensitive to minimal changes in a very limited auditory system.

One innovative approach to this problem has been developed by Owens and his co-workers (Owens, Kessler, & Schubert, 1982; Owens, Kessler, Telleen, & Schubert, 1981). Their Minimum Auditory Capabilities MAC test was specifically designed to determine whether profoundly hearing-impaired subjects can perform such speech processing tasks as inflection (question vs. statement), accented word, and closed-message-set word identification. Results reported to date suggest that even these extremely restricted listening tasks may be too difficult for many profoundly hearing-impaired subjects.

It is important to note that almost everyone who has worked with implanted patients reports improvement in lipreading skills as a possible reason for successful rehabilitation. This suggests the need to expand the concept of speech recognition, at least in this subgroup of hearing-impaired patients, to include all aspects of human communication. Not only lipreading but also context, speaker familiarity, and emotional loading of the message are potential factors in speech recognition of the profoundly hearing-impaired.

It may be that the wrong question is being asked in speech reception testing of the profoundly hearing-impaired. Rather than designing a variety of tests to differentiate various aspects of speech communication, perhaps a test based on the concept of a hierarchy of extraauditory cues to the recognition of a spoken message should be considered. By such an approach, the same message would be presented to the subject repetitively, with a different extraauditory cue added to each repetition. The level of extraauditory cue required for successful speech communication would then be the measure of interest.

At the present time, research in this aspect of speech reception testing will result in direct benefit to only a limited subgroup of patients. This is a relatively new application, brought about by advances in medical technology, and could result in a fresh approach to the application of speech reception testing in the evaluation of social adequacy.

The final diagnostic application of speech reception measures, evaluation of performance with a hearing aid, is probably the most controversial. Previous investigations of this application have revealed the unreliability of many speech measures for this purpose. Many investigators eschew speech measures in hearing aid evaluation and approach the fitting problem by selective amplification. In fact, hearing aid fitting on the basis of spectral consideration is probably entirely appropriate in patients with cochlear dysfunction. It is less appropriate, however, when the auditory disability is the combined result of peripheral and central effects. This emphasizes, again, the research needs in speech reception evaluation of the elderly—the single largest subgroup of potential hearing aid users.

At the present time there is little justification for formal speech reception comparisons of aided performance. In the absence of evidence showing that either unaided or aided performance predicts successful use of amplification, formal hearing aid evaluation measures are a waste of valuable clinical time. An important goal for future research is to validate all aspects of aided and unaided speech reception performance as it relates to successful use of a hearing aid.

The real challenges in this area include (a) identification of an appropriate index of successful hearing aid use; (b) control for the usual host of hearing aid variables such as gain setting, ear selection, output limitation, earmold, and so on; and most importantly, (c) selection of appropriate speech materials for unaided and aided comparisons. In a recent study at Baylor which surveyed hearing aid use and satisfaction, significant soundfield aided performance differences between satisfied and dissatisfied hearing aid users were observed. As a result, it is believed that a key guideline for future research is the use of a speech task whose degree of listening difficulty can be manipulated over a wide performance range.

In summary, the diagnostic applications of speech reception testing were defined as any measure used to identify, evaluate, or rehabilitate the hearing-impaired patient. The scope of diagnostic speech reception measures has changed considerably in the past 30 years. Emphasis has shifted from previous applications, identification or confirmation of sensitivity loss, and localization of site of auditory disorder, to aspects important for evaluation of rehabilitation potential. There are three important areas of future research in diagnostic speech reception measures: (a) differentiation of peripheral versus central effects in the elderly, (b) evaluation of the profoundly hearing-impaired, and (c) prediction of potential for success of hearing-aid fitting.

REFERENCES

OWENS, E., KESSLER, D., & SCHUBERT, E. D. (1982). Interim assessment of candidates for cochlear implants. *Archives of Otolaryngology, 108*, 478–483.

OWENS, E., KESSLER, D., TELLEEN, C., & SCHUBERT, E. D. (1981, September). The minimal auditory capabilities (MAC) battery. *Hearing Aid Journal, 34*, 9–34.

Chapter 10

LINGUISTIC STRUCTURE IN CLINICAL AND EXPERIMENTAL TESTS OF SPEECH RECOGNITION

JOHN J. GODFREY

Callier Center for Communication Disorders
University of Texas at Dallas
and
Institute for Defense Analyses, Princeton, NJ

The following is a brief overview, both selective and subjective, of a rather broad subject, divided roughly into two areas. The first part, which is mainly retrospective, reviews some traditional questions on the linguistic structure of speech stimuli of the isolated word or syllable type in clinical and research tests. The second, mainly prospective, questions the further relevance of this enterprise, and looks briefly at the Speech Perception in Noise (SPIN) Test as a potential point of departure for future tests and experimental paradigms which might incorporate some of the insights of recent speech and language research.

Linguistic contribution to speech audiometry from the 1940s through the 1960s and even beyond have come principally from phonetics in the design and refinement of word lists used for articulation testing and its audiometric analogues, speech reception and speech discrimination testing (Davis & Silverman, 1970; Egan, 1948; Fletcher, 1953; Goetzinger, 1972). Although there is no need to review in much detail the forms these contributions took, it should be noted at least in passing that the concept of the phoneme played a central role in this transfer of information from linguistics to audiology.

When the early word lists were designed for their several purposes, one of the test design requirements was that the stimulus materials be representative of everyday speech, presumably to ensure that they constituted in some sense a random sample of the domain about which inferences would be made. The interpretation given this requirement was that individual speech sound segments, or phonemes, should occur with approximately the same relative frequency as in spoken English (or in some class of written text), which came to be known as *phonetic balance* (PB), or more accurately, as Lehiste and Peterson (1959) pointed out, *phonemic balance.* This PB condition was imposed within other constraints; typically, the materials were to be meaningful (rather than nonsense items), of low redundancy (single words rather than phrases or sentences), and usually limited further to words of one syllable or two stressed syllables and to words of a certain presumed familiarity (most often satisfied by statistics of frequency of occurrence). Even granting the assumptions of phonemic balancing, these constraints introduce a number of biases: Monosyllabic words restrict the types of consonant sequences and allophones which may appear; the intervocalic flapped /r/, for example, as in *writer*, *rider*, *better*, may not occur in monosyllables, nor in so-called "spondees," though its textual frequency in English is high.

If vowels are varied, monosyllables also tend to be very dialect sensitive; many studies using such items, including the classic Peterson and Barney experiment (1952), probably have artificially high error rates because some of their listeners distinguish the vowels in *tot/taught*, or in *pin/pen*, whereas others do not. Moreover, many of the important dynamic cues in speech, such as the relative durations of vowels, glides, and fricatives, are normally evaluated with respect to both the tempo (Fitch, 1981) and the syntactic structure (Klatt, 1975) of an entire utterance. The list of such limitations could be made very long, but they simply amount to the observation that the phonetic structure of a list of short words can only be made to reflect the phonology of short words.

Within these limitations, the composition of stimulus materials for speech recognition testing progressed in a number of ways, with refinements or alternative materials being developed as research results showed the need. Thus, for example, the range of word familiarity was more strictly controlled in the CID test lists (Hirsh et al., 1952). Lehiste and Peterson (1959) produced word lists in which syllable type was also controlled; instead of allowing both open and closed syllables to compete in a list, they used only closed syllables (CVC). This permits the maximum number of vowel distinctions and yields the largest number of words of identical syllable structure. Their lists also were designed to conform to a stricter definition of phonemic balance, which took account of the different distributional properties of English consonantal sounds in the prevocalic and postvocalic positions of stressed monosyllables. In order to make their lists adaptable for clinical as well as research purposes, Peterson and Lehiste (1962) later revised them by eliminating less familiar words.

Another approach was to focus on what might be called individual *phoneme reception*, thus bypassing the need for phonemic balance throughout the entire list. The Rhyme Test (Fairbanks, 1958) did this by confining the possible locus of errors to one segment of the stimulus. The closed response format of the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965) further limited the number of phonemic alternatives available to the listener for each stimulus item in order to eliminate response variability based on the individual's vocabulary, the number of competing items in the language, and the interaction of these factors. The use of carrier phrases was also widely adopted, especially in the clinic, as a means of adding at least some form of context to the isolated test words. It should be noted, however, that this is only minimally achieved when the identical carrier phrase is repeated and when metalinguistic reference is made to the test word, as in "Say the word ______."

In addition to the clinical use of word lists, wherein speech discrimination ability is measured in the number or percentage of items correctly identified, a number of research studies have used item analysis of errors or stimulus-response patterns to improve our understanding of the process of speech perception either by impaired ears or in conditions simulating impairment (Bilger & Wang, 1976; Danhauer & Singh, 1975; Miller & Nicely, 1955; Owens, Benedict, & Schubert, 1972; Sher & Owens, 1974; Wang & Bilger, 1973; Wang, Reed, & Bilger, 1978). Some use simple words, though most use nonsense syllables as stimuli in order to avoid redundancies provided by listeners' knowledge of the lexicon and thus force only errors due to misperception of acoustic cues. Since the focus in these studies is most often on specific effects of peripheral hearing loss, noise, or filtering on the perception of speech sounds, rather than on speech understanding or social adequacy, and since it is possible to train or to select subjects for the tasks involved, this restriction is appropriate. But no one should jump to the conclusion that a neat separation of "levels of processing" is thereby achieved, as for example between speech perception and language comprehension. Recent psycholinguistic work suggests that this may not be achievable in any useful sense.

Insofar as the errors or confusions generated by these experiments are examined in terms of the acoustic properties involved, they provide valuable knowledge about the effects of sensorineural impairment on speech perception. For example, because sibilants as opposed to other fricatives are nearly always signaled by relatively intense high-frequency acoustic energy, and nasal consonants by a broad low-frequency band, the nasal sounds tend to be missed less and the sibilants more under conditions of sloping loss or low-pass filtering. However, if an analysis is made only, or primarily, in terms of an a priori set of linguistic categories (phonemes or distinctive features), the results are likely to be less enlightening. This is just what Wang and Bilger (1973) found in one of the most thorough studies of this type. Even with a carefully restricted set of nonsense monosyllabic stimuli, given a choice among several proposed distinctive-feature nomenclatures with which to rationalize the results, none proved outstanding as models of the perceptual dimensions of their subjects' performance. This result is not puzzling or contradictory, but quite natural; in what follows, some reasons are given for why linguistically defined structural units may fail to predict perceptual effects in such experiments, at least in enough detail to satisfy the requirements of hearing science. The first reason, which applies to both phonemic and distinctive feature analysis, has to do with the nature of the relationship between linguistic message units and the acoustic signal; the second involves the larger question of how entire utterances are represented—that is, which language model is correct.

In pretransformational linguistic terms (cf., for example, Jakobson & Halle, 1956), the *phoneme* is generally understood to mean a minimal sound segment which fulfills a *distinctive function* in a given language—that is, serves to differentiate words or morphemes with different meanings. The term *segment*, however, has a peculiar meaning in this definition; to the linguist at least, it refers not to a physical portion of a *signal* (an observable articulatory or acoustic output), but to a position in its left-to-right linguistic representation (the *message*), which is logically prior to its physical realization. Thus, phonemes are commutable and permutable in a sense that their corresponding physical signals are not. Moreover, if we think of a phoneme as a selection from among a fixed number of alternatives for a given position, it is not hard to demonstrate that this selection is better regarded as a set of simultaneous (perhaps binary) choices, which are the *distinctive features*. What is observed in the signal are the articulatory/acoustic consequences of these sets of choices, as they enter into the multitude of simultaneous and sequential combinations permitted by the language.

The point of reviewing these distinctions is to emphasize that the relationship between message categories and signal properties is not always straightforward. For example, in the monosyllables *yes* and *say*, the phonemic representations /jɛs/ and /se/ have a different number of phonemes and different vowels, so that for purposes of comparison they are /$C_1V_1C_2$/ and /C_2V_2/, respectively. Yet the acoustic signal will seem to belie this difference—If a typical token of one is played backwards, it will be heard as the other. To the linguist, the vowel and the diphthong are commutable, as in *say/saw/see*, and so on, whereas the initial glide is commutable with other initial consonants as in *yes/guess/Wes/less*, and so on. To the auditory system the frequencies, durations, and rates of change are about the same in tokens of either. A pair of monosyllables like *trot* /trɑt/ and *tart/* tɑrt/, will present the opposite problem. The acoustic realizations of the /r/ will differ in a number of ways, the most important being that the /r/ in *trot* (phonetically [tɽɑt]) will have an aspirated (voiceless) source, hence relatively little energy below 1 kHz. In fact, it is perhaps most likely in this context to be confused with the voiceless fricative /ʃ/ (cp. *true/chew*, phonetically /tʰɽu/-/tʃu/), which is a highly unlikely substitution in other contexts for a voiced sonorant or glide.

Such examples merely show the well-known dependence of acoustic cues on local phonetic context, expressed in terms of phonemic symbols; feature notation could just as well have been used. Similar interactions are found among distinctive features within a single segment. For example, the *voicing* feature is realized in quite different ways when combined with the feature [+ continuant], or *fricative*, than with [− continuant], or *stop*. This difference is hardly trivial where characteristics of hearing loss are at issue, since one cue is principally spectral and the other temporal, at least in initial prestressed position. In word-final position still other differences prevail, with reciprocal length changes in consonant and vowel being highly audible cues to voicing for both stops and fricatives. Thus, it is not surprising that monosyllabic items distinguished by voicing are well preserved as such under certain conditions of hearing loss or noise, but it is not necessarily due only, or even principally, to the low-frequency content of the segments marked [+ voice]. Especially if final consonants are present, it may be due in some measure to the preservation of timing and duration cues. In fact, there is evidence to this effect both for normal listeners (Hogan & Rozsypal, 1980; Raphael, 1972) and for hearing impaired (Revoile, Pickett, Holden, & Talkin, 1980). Likewise, the cues to the [+ nasal] character of a syllable-final nasal consonant are perhaps more strongly represented in the preceding vowel, in the sense of being carried by the more intense vocalic portion of the signal, and across much of its spectrum, than in the weaker "murmur" portion. This might result in more errors for /n/ in *nap* than in *pan* for example, although the feature specifications would be the same.

The relationship between distinctive features and acoustic cues is often complex. One may well ask whether there is any sense in which the characterization of a simple English word in terms of a matrix of distinctive features can be said to predict its acoustic realization. The answer, according to Stevens and others, is yes, but in a rather restricted sense (Stevens, 1980; Stevens & Blumstein, 1980). Perhaps the clearest explanation of this viewpoint, which might be called the "relational invariance" hypothesis, is found in chapters 9 and 10 of Fant's *Speech Sounds and Features* (1973). Briefly, it holds that, at least in careful pronunciation, a given distinctive feature will always be realized by some articulatory/acoustic property which will be invariant in a relational sense, that is, measurably distinct from the value that its polar opposite would have had in the same context. This is still a controversial hypothesis; however, even if true, it does not imply that that particular property will be the most prominent cue or the one used by the ear to perceive the distinction in that context, much less under conditions such as noise or hearing loss. Other cues, such as vowel length or nasality in the examples above, may be more effective, but are by definition "secondary" or "redundant" with respect to the bipolar (binary) distinction in question.

A further complication in the contemporary use of the term *distinctive feature* with reference to auditory data is that, within the context of transformational linguistics, features also function as constants in derivational formulas, where they must satisfy a number of formal constraints quite foreign to the domain of hearing science. The number and names of distinctive features proposed at any given time is therefore largely dependent on knowledge from very different sources (Halle, 1977a). At one time, for example, Halle (1977b) proposed to distinguish American English vowels as both tense/lax and long/short, with 18 "underlying" simple vowel nuclei. Such a system could not be arrived at by acoustic phonetic analysis alone (since some of the underlying vowels never occur in pronunciation), but was proposed mainly because (a) the length feature was required to derive stress by a certain set of rules and (b) tenseness to derive certain alternating forms such as *lie-lay*, *write-wrote*, and so on, by the Vowel Shift Rule. Choosing among sets of distinctive features motivated in part by this type of evidence, if one's hypothesis relates to sound structure only, would be of doubtful value in the present state of our knowledge. In short, linguistically defined phonological units are still rather coarse instruments with which to analyze perceptual responses to the acoustic structure of speech (Fant, 1973, pp. 162–165), given the kind of detail that is of interest to hearing science.

Fortunately, a great deal of detailed knowledge has been accumulated about the nature of the acoustic cues to speech. Frequently, one is able to specify the characteristics of experimental speech stimuli in terms of their acoustic and even auditory properties, to synthesize them for experimental purposes, and even to relate specific properties to psychoacoustic variables. This is a productive way to organize research hypotheses about particular speech cues and impaired hearing, especially given the prospect of the application of signal-processing technology to hearing-aids, since the signals of greatest interest are obviously those of speech.

The preceding sections have dealt primarily with what is commonly called, in the jargon of speech recognition, "bottom-up," as opposed to "top-down," processes. In providing quantifiable means of relating the comprehension of speech to objective properties of sound, listening channel, and characteristics of hearing, there is no doubt that both the clinical tests and the research studies employing controlled sets of speech stimuli have contributed significantly to understanding speech perception, both normal and impaired, and that the stimulus sets were much more efficient for having been constructed with these factors taken into consideration. In applying such tests to clinical populations, however, the prior assumptions of articulation testing were not often seriously questioned, "the difference in procedure being simply that the listener became the variable while the speaker and transmission system were controlled" (Owens, 1961, p. 113). They also fit quite comfortably with assumptions once current in American Structuralist linguistics, notably the separability of meaning and context from the "real" structure of language, which was in turn divisible into distinct "levels" of syntax, morphology, and phonology.

Although these views have long been extinct, especially as applied to real linguistic behavior, there has been little consequent effort to revise the view of linguistic behavior in the hearing impaired. Audiologists continue to count missed phonemes, while they also puzzle over the inability of standard speech test results to predict everyday communicative performance. In the absence of much experimental evidence, perhaps some speculation on this topic is not out of order. Psycholinguistic research has shown repeatedly that top-down and bottom-up processing interact in numerous complex ways, so that any change of acoustic input, or any change in available interpretive information, may have effects at any level. To choose just one example, the research of Marslen-Wilson and others (Marslen-Wilson, 1975; Marslen-Wilson & Welsh, 1978; and references therein) on the shadowing of meaningful spoken passages shows that subjects who are repeating speech less than a quarter of a second behind the model will, without detecting it, correct a polysyllabic mispronounced word into the correct word based on the first syllable heard. Even if these latencies are lower limits for practiced subjects, the implication for real-life situations still seems clear: Contextual (semantic, syntactic, lexical) constraints sufficient to predict the rest of the word are continuously being brought to bear on the acoustic input at the earliest possible stage. Much of the recent literature on *lexical access* is continuous speech, in spite of disagreement on details, supports this view of intimate and parallel interaction between top-down and bottom-up searching (Cole, 1980; Levelt, 1978; Morton, 1979).

Consider another, seemingly unrelated datum: Pisoni (1981) showed that LPC-synthesized approximations to real speech, even when 100% intelligible, cause significantly longer reaction times in lexical decision tasks. This situation, in which an inferior acoustic input causes perceptual difficulty not directly reflected in missed test items, is reminiscent of the clinical situation we are interested in.

Taken together, such results suggest a *resource allocation* model for problems of speech understanding with hearing impairment, which can hardly be made to fit the speech recognition test situation. These results imply that all lost acoustic information has its price, not necessarily measurable in feature counts or specific misidentifications, but at times in effects as subtle as less time and/or information available for decisions about the tense of a verb or the intended reference of a pronoun.

Effects such as these cannot be observed, much less measured, except in contexts much closer to real communicative situations. Complete grammatical and meaningful sentences are certainly the minimum, and the Speech Perception in Noise (SPIN) Test (Kalikow, Stevens, & Elliott, 1977) is an important step in the right direction. If suggestions for improvements in the linguistic structure of the SPIN test materials were considered, they should not be at the phonetic balance of the test words, but at such factors as homogeneity of syntactic structure (e.g., eliminating or controlling the current mix of simple and complex sentences) and communicative import (e.g., providing more realistic or natural-sounding carrier sentences for the low-predictability words). The basic design of the test and the normative data gathered with it should inspire some experimental studies which vary these and other factors that are either fixed or randomized in the original, such as the syntactic function and prosodic status of the test word.

Beyond its clinical value, once standardized and perhaps improved in a few respects, this test format will stimulate ideas for a new generation of research investigations along the lines suggested above, opening up broader psycholinguistic questions than were possible in the era of counting missed phonemes. One can only hope that this will lead in turn to better instruments for predicting real communicative performance and to better means of improving it for individuals with impaired hearing.

REFERENCES

BILGER, R. C., & WANG, M. D. (1976). Consonant confusions in patients with sensorineural hearing loss. *Journal of Speech and Hearing Research, 19,* 718–748.

COLE, R. A. (Ed.). (1980). *Perception and production of fluent speech.* Hillsdale, NJ: Erlbaum.

DANHAUER, J. L., & SINGH, S. (1975). *Multidimensional speech perception by the hearing impaired.* Baltimore: University Park Press.

DAVIS, H., & SILVERMAN, S. R. (Eds.). (1970). *Hearing and deafness* (3rd ed.). New York: Holt, Rinehart & Winston.

EGAN, J. P. (1948). Articulation testing methods. *Laryngoscope, 58,* 955–991.

FAIRBANKS, G. (1958). Test of phonemic differentiation: The Rhyme Test. *Journal of the Acoustical Society of America, 30,* 596–600.

FANT, G. (1973). *Speech sounds and features.* Cambridge: MIT Press.

FITCH, H. (1981). Distinguishing temporal information for speaking rate from temporal information for intervocalic stop voicing. *Haskins Laboratories Status Reports, SR-65,* 1–32.

FLETCHER, H. (1953). *Speech and hearing in communication.* New York: van Nostrand.

GOETZINGER, C. (1972). Word discrimination testing. In J. Katz (Ed.), *Handbook of clinical audiology.* Baltimore: Williams & Wilkins.

HALLE, M. (1977a). Tenseness, vowel shift, and the phonology of the back vowels in modern English. *Linguistic Inquiry, 8,* 611–625.

HALLE, M. (1977b). [Review of S. Singh, *Distinctive features: Theory and validation*]. *Journal of the Acoustical Society of America, 62,* 801–802.

HIRSH, I. J., DAVIS, H., SILVERMAN, S. R., REYNOLDS, E., ELDERT, E., & BENSEN, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17,* 321–337.

HOGAN, J. T., & ROZSYPAL, A. J. (1980). Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *Journal of the Acoustical Society of America, 67,* 1764–1771.

HOUSE, A. S., WILLIAMS, C. E., HECKER, M. H. L., & KRYTER, K. D. (1965). Articulation testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37,* 158–166.

JAKOBSON, R., & HALLE, M. (1956). *Fundamentals of language.* The Hague: Mouton.

KALIKOW, D. N., STEVENS, K. N., & ELLIOTT, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61,* 1337–1351.

KLATT, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3,* 129-140.

LEHISTE, I., & PETERSON, G.E. (1959). Linguistic considerations in the study of speech intelligibility. *Journal of the Acoustical Society of America, 31,* 280–286.

LEVELT, W. J. (1978). A survey of studies in sentence perception: 1970-1976. In W. J. Levelt & G. M. Flores d'Arcais (Eds.), *Studies in the perception of language.* New York: Wiley.

MARSLAN-WILSON, W. D. (1975). Sentence perception as an interactive parallel process. *Science, 189,* 226-228.

MARSLAN-WILSON, W. D., & WELSH, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29-63.

MILLER, G. A., & NICELY, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

MORTON, J. (1979). Word recognition. In J. C. Marshall & J. Morton (Eds.), *Psycholinguistics 2: Structures and processes.* Cambridge: MIT Press.

OWENS, E. (1961). Intelligibility of words varying in familiarity. *Journal of Speech and Hearing Research, 4,* 113–129.

OWENS, E., BENEDICT, M., & SCHUBERT, E. D. (1972). Consonant phonemic errors associated with pure-tone configurations and certain kinds of hearing impairments. *Journal of Speech and Hearing Research, 15,* 308–322.

PETERSON, G. E., & BARNEY, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

PETERSON, G. E., & LEHISTE, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Research, 27,* 62–70.

PISONI, D. B. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America, 70*(Suppl. 1), S98.

RAPHAEL, L. J. (1972). Preceding vowel duration as a cue to the voicing characteristic of the word-final consonants in American English. *Journal of the Acoustical Society of America, 51,* 1296–1303.

REVOILE, S., PICKETT, J. M., HOLDEN, L., & TALKIN, D. (1980). Effects of some acoustic cue modifications on the perception of voiced and unvoiced final stop consonants for hearing-impaired listeners. *Journal of the Acoustical Society of America, 67*(Suppl. 1), S78.

SHER, A. E., & OWENS, E. (1974). Consonant confusions associated with loss above 2000 Hz. *Journal of Speech and Hearing Research, 17*, 669–681.

STEVENS, K. N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America, 68*, 836–842.

STEVENS, K. N., & BLUMSTEIN, S. E. (1980). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 1–38). Hillsdale, NJ: Erlbaum.

WANG, M. D., & BILGER, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America, 54*, 1248–1266.

WANG, M. D., REED, C. M., & BILGER, R. C. (1978). A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusion. *Journal of Speech and Hearing Research, 21*, 5–36.

Chapter 11

PROCEDURAL FACTORS IN SPEECH RECOGNITION TESTING

HARRY LEVITT

City University of New York, New York

The purpose of this paper is to identify the basic principles underlying the test procedures used in speech recognition testing. These principles are quite general and apply to all forms of testing. The examples used to illustrate these principles, however, are restricted to the problems of speech recognition testing.

The most pervasive principle of all is that uncertainty has two basic forms, uncertainty that can be quantified and uncertainty that simply cannot be quantified. These two forms of uncertainty take on different guises, depending on one's vantage point (Levitt, 1983). In the world of testing, *quantifiable uncertainty* manifests itself as lack of precision and *unquantifiable uncertainty* as lack of accuracy.

The distinction between precision and accuracy is extremely important, yet often misunderstood. Precision is essentially a measure of repeatability. If a measure is repeated many times and the measurements do not differ by very much, then they are said to be precise. If the measurements differ substantially then they are said to be imprecise. Degree of precision can be specified in many ways. A common method is to specify precision of measurement in terms of the standard deviation of the measurements.

Accuracy is essentially the difference between that which is actually measured and that which is believed to be measured. Typically, the accuracy of a measurement is defined as the difference between the true value of the quantity to be measured and its precisely measured value. Since the true value of an unknown quantity can never be determined with certainty, accuracy of measurement is not quantifiable in absolute terms. It is possible, however to specify relative accuracy, for example, by comparing a new or approximate method of measurement with an accepted standard procedure.

The inherent limitation in precision of measurement is the random variability of the measurements. Precision can be improved by taking the average of many measurements. Several basic theorems in statistics specify the extent to which precision is improved with an increase in the number of measurements. A fundamental consideration in these theorems is that the uncertainty associated with lack of precision can be quantified.

The inherent limitation in accuracy of measurement is the validity of the underlying assumptions. In order to measure, it is necessary to have some concept or model of what is being measured. For example, one would not attempt to measure the current in an electric circuit with a ruler. The concept of that which is desired to be measured is in essence a model of the quantity to be measured.

The above point is perhaps best illustrated by a comparison between subjective measurement, such as speech recognition testing, and physical measurement, such as the measurement of electrical current. Although the exact nature of an electrical current is not known, there is a well-developed theoretical model of electricity that allows for extremely precise measurements of such quantities as current, voltage, and resistance. There is also a remarkably high degree of consistency between these precise measurements and the underlying model of electricity. It is nevertheless possible that the concept of electricity may be quite wrong, in which case the existing body of measurements of electrical quantities would be inaccurate. Since the accuracy of the assumptions underlying the commonly accepted model of electricity cannot be verified in absolute terms, the accuracy of electrical measurements cannot be quantified absolutely.

It is widely believed that physical measurements are inherently more accurate than subjective measurements. This is not necessarily true, because the model underlying either type of measurement can be in error by an unknown amount. The important difference between physical and subjective measurement is the high precision and degree of consistency with underlying models that have been obtained with physical measurements in comparison with subjective measurements. In particular, precision of measurement in speech recognition testing is relatively poor, as is the degree of consistency between theories of speech perception and associated measurements. Nevertheless, although there are vast differences between physical and subjective measurements in terms of relative precision and degree of consistency, the same general principles apply to both sets of measurements.

An important practical consequence of the relatively poor precision and internal consistency of speech recognition measurement is that models of human speech recognition are constantly being revised in an attempt to improve their validity. This creates serious problems of interpretation because there has yet to be developed a generally accepted standard method of speech recognition testing against which other methods of measurement or new theories can be compared and evaluated.

A second general principle is that the more that is known about the quantities to be measured, the more effectively these quantities can be measured. This creates a "catch-22" situation in that in order to measure a quantity precisely and accurately it is necessary to know in advance that which is to be measured. The usual situation in practice is that in which limited information is available on the quantity to be measured and an experiment is performed in order to obtain more information on the quantity of interest. It is possible to design an experiment so as to maximize the gain in information. This is normally done in terms of increasing the precision of measurement since theoretical treatments of this problem require some way of quantifying the gain in information and, as noted earlier, only one form of uncertainty (lack of precision) can be quantified.

A very powerful method of measurement is that in which the experiment is redesigned after every observation so as to maximize the gain in information on the next observation. In this way, all of the available information is used at any given time to maximize the precision of the next measurement. This approach to testing is known as adaptive testing, because the method of measurement adapts itself to the data being generated. In statistics, efficiency is usually measured in terms of the precision per observation. Adaptive procedures can thus be designed to be extremely efficient. Although adaptive testing can be very complicated, it is possible to develop relatively simple methods of adaptive testing that are almost as efficient as the most sophisticated methods.

An example of a simple, relatively efficient adaptive technique is the simple up-down procedure (Dixon & Mood, 1948) and variations of it, such as the transformed up-down procedure (Levitt, 1971; Wetherill & Levitt, 1965). These techniques are discussed in some detail because they provide good illustrative examples of the principles being discussed and they are also frequently used in speech recognition testing.

The simple up-down procedure is often used in obtaining the 50% level of the performance-intensity function (also known as a response curve or psychometric function). Figure 1 shows a typical performance-intensity function and Figure 2 shows a typical sequence of trials using the simple up-down procedure. An arbitrary scale is used in specifying the stimulus levels. In practice, the stimulus scale may be dB SPL, signal/noise ratio in dB, or some physical measure of a degraded speech signal. The specific scale that is used is not critical to the discussion that follows, provided the response curve increases monotonically with increasing stimulus level.

The stimulus level used on the first trial is known as the initial value. After each trial the stimulus level is either increased or decreased by a fixed amount (the step size) depending on the subject's response. A positive response (e.g., correct identification of the test item, which could be a word or nonsense syllable) results in a decrease in stimulus level on the next trial. A negative response (e.g., incorrect identification of the test item) results in an increase in stimulus level on the next trial. A run consists of a sequence of changes in stimulus level in one direction only. For the example shown, Trials 1–3 constitute the first run, Trials 3–7 the second run, and so on.

The method of estimating the 50% level is related to the rule for controlling stimulus level. Both are concerned with finding the 50% level. The objectives, however, are different. The first is concerned with placing observations at or very close to the 50% level. The second is concerned with deriving an estimate of the 50% point from the data available. It is possible, but not necessary, to use the same estimation procedure in placing observations as in analyzing the resulting data. Typically, in up-down testing a very simple rule is used to place observations in the region of interest followed by a more sophisticated rule for estimating the parameters of interest.

The 50% level of the performance-intensity function can be estimated in many different ways. A very simple and relatively efficient method is to take the average of the midpoints of every second run (Wetherill & Levitt, 1965). A second simple method of estimation is to take the average of all of the

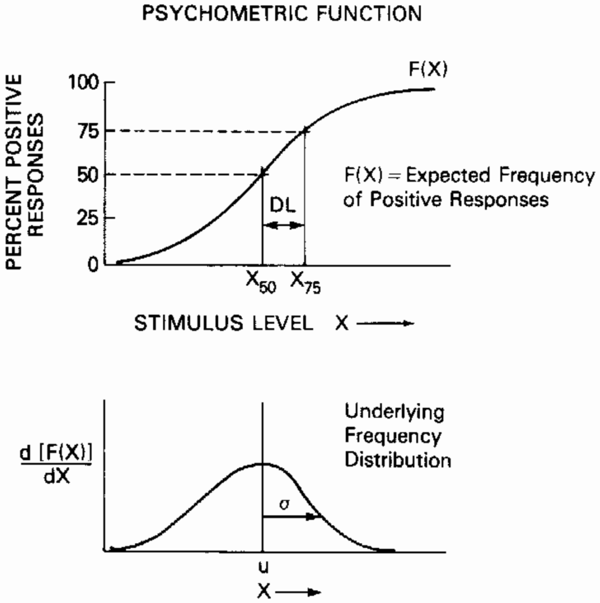

FIGURE 1. Typical psychometric function. The upper curve shows the expected frequency of positive responses in a typical experiment. In some applications, the curve may be the cumulative form of the frequency distribution shown in the lower portion of the figure (from Levitt, 1971).

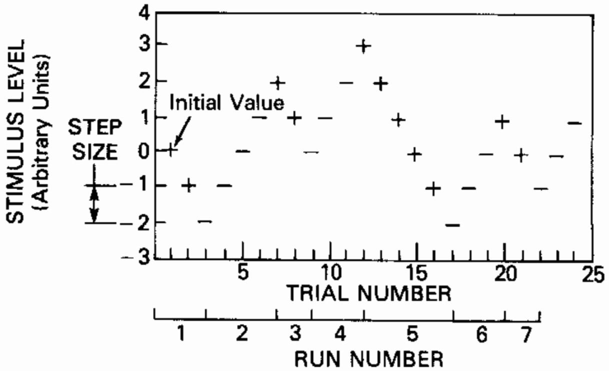

FIGURE 2. Typical data for simple up-down procedure (from Levitt, 1971).

stimulus levels used after the first run. A third, more complex method of analysis is to fit a curve to the data using an efficient method of parameter estimation (e.g., the method of maximum likelihood) and then to determine the 50% level of the fitted curve.

It can be shown theoretically that whichever method of analysis is used, in order to estimate the 50% level precisely, observations should be placed as close to the 50% level as possible. The up-down procedure has the very important property that it converges on the 50% level and that, as a consequence, observations are placed at or near the target level. An extremely precise up-down procedure would be one in which the initial value is at the 50% level, and an extremely small step size is used so that the excursions about the 50% level using the up-down rule are relatively small. If a very small step size is used, then it is very important that the initial value be at or close to the 50% level, otherwise many observations will be used inefficiently as the procedure gradually converges on the target level. Thus, in order to estimate the 50% level most efficiently, it is necessary to know the value of the 50% level in advance. It is necessary to know the overall spread of the response curve so that a reasonable step size is chosen. A step size equal to or greater than the transition of the response curve is typically too large, and one that is a minute fraction of this transition region is typically too small for practical purposes. In short, the more we know about the quantity to be measured, the more efficiently can we measure it.

Another illustrative example deals with a much more difficult problem, commonly encountered but never resolved satisfactorily—that of locating the maximum or peak of a performance-intensity function. This point is often referred to as PB-max. Note that this point has two coordinates, its location (the signal intensity corresponding to the maximum) and its value (the test score obtained at this signal intensity). The traditional method of measuring PB-max is to present an entire word list at each of several levels, chosen so as to cover the range of the performance-intensity function. The signal level yielding the highest score is used as an estimate of the location of PB-max and the test score at this level is used as an estimate of the value of PB-max. This approach has many shortcomings. Figure 3 shows two typical sets of data. In each case, the mean-squared deviation between the data and the fitted curves is not significantly greater than that to be expected from the test-retest variability of the measurements. Since there is no a priori reason for favoring any one of the fitted curves, the estimated location and value of PB-max are thus only very rough estimates.

One of the reasons for the poor performance of this method of measuring PB-max is that very little use has been made of prior information about PB-max. For example, the data points corresponding to the three lowest signal levels provide very little information on either the location or value of PB-max, That is, half the data obtained in the experiment provide very little information relative to the quantity of interest, PB-max. These three data points do provide information on the shape of the performance-intensity function at low intensity levels, but this is of little interest in the present context.

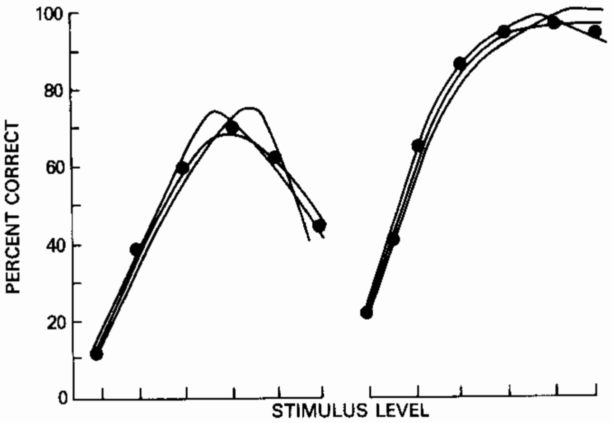

FIGURE 3. Estimating PB-max. Two sets of data are shown. Three possible curves have been fitted to each set of data.

A more efficient way of measuring PB-max is to place observations at or close to PB-max with sufficient observations on either side of PB-max to confirm the location of PB-max. If PB-max occurs at or beyond the highest signal intensity that can be used safely, then signal intensities on only the lower side of PB-max should be used (i.e., ethical concerns take precedence over statistical considerations). An adaptive procedure for estimating PB-max has been developed (Levitt, 1978) which makes increasing use of the information on PB-max as the test progresses. The technique works well when the performance-intensity function rolls over markedly; that is, the curve has a distinct peak (Kamm, Morgan, & Dirks, 1983). It does not work as well when the performance-intensity function has a large plateau. For example, the test score may increase monotonically with signal level, approaching PB-max asymptomatically at the highest signal intensities that can be used safely. For the latter case, it would be more efficient to measure loudness discomfort level directly and then check that a decrease in signal intensity produces a decrease in performance.

The use of prior information need not be restricted to information gathered within a single experiment or set of experiments. In the case of measuring PB-max, it is known that rollover occurs predominantly with certain types of impairment (e.g., retrocochlear) or under certain conditions (e.g., with intense background noise). The decision as to which test strategy to use for greatest efficiency thus depends on what is known about the subject and the likely shape of the performance-intensity function. Here again the same principle holds, the more that is known about the quantity to be measured, the more efficiently it can be measured.

The third principle is a matter of economics. Simply stated, there is a cost to everything. The process of measurement has many facets and, as a consequence, the costs involved are many and diverse. Further, costs are measured in more than one dimension and often times a compromise needs to be made between costs of quite different types. For example, poor precision, errors due to unreliable assumptions, lack of generality, excessive time and effort required of the subject, and the need for complex or expensive equipment are important costs to be considered. It is not possible to measure all of these costs on a single dimension and some insight is required of the experimenter in choosing an experimental procedure that will yield a reasonable balance of costs for the experiment at hand. Some guidelines for facilitating this decision are presented shortly.

An important consequence of the multidimensional nature of the costs involved in measurement is that there is no single "best experiment"; that is, there is no ideal test material, or test format, or experimental design, or method of analysis that is uniformly better than any other. It is possible to design an experiment that will minimize a specific cost (e.g., the number of observations needed for a specific degree of precision) but it is essential to remember that this experiment has been optimized with respect to only one of many possible criteria. It is also likely that optimization with respect to one criterion will increase costs, often substantially, with respect to other criteria. Consider, for example, the problem of estimating the 50% level of the performance-intensity function. As noted earlier, a very precise estimate can be obtained using the simple up-down procedure with a relatively small step size, provided the initial value is close to the target level. If this condition can be met, then the technique can be made to approach the maximum precision possible by having the step size approach zero. This is, however, a highly impractical way of maximizing precision. Firstly, there is only a marginal increase in precision in going from a reasonable step size (e.g., a step size roughly equal to the difference between the 50% and 70% stimulus levels) to an extremely small step size. Secondly, this limited gain in precision is achieved at the cost of an excessive dependence on the requirement that the initial value be at or very close to the target level.

Even if a satisfactory compromise is found between the conflicting costs of maximizing precision versus excessive dependence on underlying assumptions, other costs can come into play. A practical way of guarding against the effects of a poor choice of an initial value is to use a large step size at the start of the experiment and to systematically reduce the step size as the experiment progresses. In this way, the benefits of a small step size are obtained for most of the experiment while at the same time the technique is not critically dependent on a good choice of an initial value. A very efficient method of adjusting step size is that proposed by Robbins and Monro (1951) in which the step size is reduced in inverse proportion to the number of trials (e.g., if the initial step size is d, then the step size on trial n should be d/n).

This procedure is both highly efficient and relatively robust, but it does increase costs on another dimension. The necessary instrumentation is more complex because very fine adjustments in step size are required on every trial. Also, as the step size decreases, greater weight is placed on the assumption that the target level does not vary with time. Given the inherent variability of human performance, this assumption is, at best, only approximately true. There are thus practical limits as to how far the step size can be reduced. In particular, the step size should not be reduced below that needed for reliable tracking of time-dependent variations in the target level of a given subject; that is, the minimum step size should be larger than the change in target level between successive trials.

A practical rule for reducing step size that approximates the efficiency and reliability of the Robbins-and-Monro procedure, but which also takes the costs of implementation into account, is to halve the step size after a prescribed number of runs have been completed—for example, after the third run; or for a longer experiment beginning with a larger initial step size, after the first, third, and seventh run; or some similar pattern (Levitt, 1971; Wetherill, 1963).

The preceding examples show that the simple up-down procedure with a simple rule for controlling step size is not optimum on any single dimension, but it has the important virtue of being close to optimum on each of several dimensions. Although it is a very efficient procedure, it is not the most efficient procedure of its type. Similarly, it is a robust procedure but not the most robust of all such procedures. It is also a relatively simple procedure to implement in practice, but not the simplest possible procedure.

The simple up-down procedure has been chosen to serve as an example for several reasons. The procedure is a simple one and hence the complexity of the issues involved are not hidden in the complexity of the technique. It is also a procedure that is widely known and often used in speech testing. Most important of all, it illustrates how diverse costs can be taken into account in a practical way.

The fourth principle is the most general and abstract of all. Specifically, general principles apply to general principles. Thus, for example, the distinction between the two forms of uncertainty places constraints on how effectively an experimenter can make use of a priori information on the quantity to be measured. A priori information can be both imprecise and inaccurate. Since degree of precision (or its lack of) can be specified quantitatively, it is possible, in principle, to develop adaptive strategies that maximize precision with each new observation. A similar treatment for maximizing accuracy per observation is not possible because accuracy per se cannot be quantified. Relative accuracy regarding a standard procedure is specifiable and methods of maximizing relative accuracy can be developed. It is important to bear in mind, however, that the validity of the assumptions underlying a given procedure, including methods of utilizing a priori information, cannot be known with certainty. It is possible to identify faulty assumptions in terms of their lack of consistency with observation, but the reverse is not true in that consistency does not guarantee accuracy. Thus, the use of a priori information in developing more effective test strategies as well as methods for assessing costs of different techniques are subject to errors of unknown magnitude depending on the accuracy of the underlying assumptions. Nothing is known for certain, but at least if assumptions and observations are mutually consistent, there is a framework on which to build.

Having identified the major principles underlying the measurement process, it is possible to develop a set of practical guidelines for choosing an appropriate test procedure. In keeping with the general principles outlined, no claim is made that these guidelines are optimum in any general sense.

The first step is to decide on the overall purpose of the experiment. A very important distinction is whether the experiment is to provide information on performance or diagnostic (i.e., analytic) information. It is possible to get both, but compromises must be made since the requirements of an analytic or diagnostic experiment typically conflict with those for obtaining performance measures. For example, a forced-choice experiment using monosyllabic test materials can be a very effective procedure for testing a specific hypothesis in speech perception or for investigating the nature of a subject's hearing impairment, but this procedure is not well suited for obtaining an overall measure of the subject's ability to communicate with others. It is possible to predict performance from analytic information but predictions are not as reliable as direct measurement of performance. Further, the method of prediction will need to be checked with measures of performance.

The second step is to make a preliminary choice of the procedures to be followed at each stage of the experiment. These include the choice of test material, test format, experimental design (e.g., adaptive vs. fixed design), and method of data analysis. Note that these stages are closely intertwined and that each cannot be evaluated independently of the others.

The third step is to identify the costs associated with the preliminary choices made and to rank them in order of importance. This is the most difficult part because, as noted earlier, these costs are not directly comparable. Also, many of these costs cannot be quantified reliably. Nevertheless, a choice must be made in order to arrive at a practical test procedure. Even if the assessment of relative costs is not addressed explicitly, the final choice of test procedure represents a choice between conflicting costs. If this choice is made blindly, the wrong costs may be emphasized.

The fourth step is to revise the preliminary procedural choices so as to reduce costs. A suggested approach is to minimize the largest cost as far as possible, then to reduce the next largest cost (without substantially increasing the primary cost) and so on. There is no simple rule for reducing costs. It is critical when making these decisions to have a good understanding of the issues involved and not to lose sight of the overall objectives of the experiment.

The above process is made more difficult by the heavy interdependence between the various stages of an experiment. The method of data analysis, for example, is dependent on the experimental design. Ideally, an efficient experiment should be followed by an efficient method of analysis. It is often the case that given the former, a relatively simple and slightly less efficient method of analysis will be found to be satisfactory. The problem to avoid, which unfortunately occurs all too often, is that of a poor experimental design followed by a highly sophisticated method of analysis designed to extract whatever useful information can be salvaged.

In the example cited earlier (see Figure 2) it makes relatively little difference which method of data analysis is used, provided the choice of initial value and step size are appropriate. On the other hand, had the experiment been badly designed (e.g., too small a step size with a very poor initial value), then the two simpler methods of analysis described earlier would yield biased and/or unreliable estimates. The third, more complicated method of analysis would then have to be used, but even if this were to be done, the resulting estimates would not be nearly as good as those obtained simply with a good experimental design.

By analogy with the above, an efficient experimental design is heavily dependent on a good choice of test format. There is no simple rule as to which test format is best because those test formats that, in theory, yield the greatest amount of information per observation are also the ones that are heavily dependent on questionable assumptions. For example, an N-alternative forced-choice format yields a good deal of information per observation when N is very large. A necessary assumption, however, is that the subject can memorize all these alternatives and reliably compare the test stimulus with each alternative. Exactly how many of these alternatives can be used reliably depends on the test material. For auditory tests in which the alternatives appear in temporal sequence, values of N as high as 4 can be used with experienced subjects (Swets, 1964). It is more common, however, to use an N of 2 or 3. For tests in which the alternatives are presented graphically, as in the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965) or the CUNY Nonsense Syllable Test (Levitt & Resnick, 1978), values of N as high as 6 or 8 have been used reliably with adults. For young children, much smaller values of N (2 or 3) are typically used.

The closed-response-set format has the important advantages of relatively high precision and small adaptation or learning effects. These advantages are achieved, however, at the cost of superimposing an artificial response format and possibly missing important aspects of the subject's perception of speech cues. For example, a hearing-impaired subject may hear sounds that are not included in the response set and as a consequence the experimenter will obtain no information on what these sounds may be. Another possibility is that in a test of the modified rhyme type in which a finite number of alternatives is provided for a consonant at the beginning or end of a word, the hearing-impaired subject may simply not hear a consonant at all. In this case, each of the available alternatives is equally unlikely and the subject will be guessing at random. Further, the subject's guesses may not be evenly distributed across the available alternatives, but rather exhibit a bias towards one or two favored alternatives. The above problems can be handled by balancing test items so as to identify biased guessing and analyzing the resulting data accordingly. The point is that no test format is without its limitations, and it is important to know of these limitations when choosing the most appropriate test format for a given set of conditions and type of subject.

Given an efficient and reliable test format, an efficient and reliable experimental design can be developed, such as a well-designed adaptive test procedure. However, if a poor test format is used, for example, a yes/no procedure with no control over the subject's choice of criterion, then relatively little advantage will be gained from a highly efficient experimental design, since only a small fraction of the information that is potentially available will be obtained per observation.

Perhaps the greatest degree of interdependence lies in the choice of test material. The test format, experimental design, and method of data analysis are all heavily dependent on the choice of test material. In order to facilitate this discussion, speech test materials will be grouped into four broad classes: nonsense syllables, single words, sentence-length materials, and continuous or interactive discourse.

Nonsense syllables are well suited for analytic testing. The parameters of each nonsense syllable item can be strictly controlled as well as the response format. These materials can be easily used in either a closed-response-set format or an open-response format. Similarly, precise methods of controlling stimulus level can be achieved using either a fixed or adaptive experimental design. Within this context, it is not unduly difficult to take into account the effects of common sources of variability such as between-talker and between-utterance differences. This can be done by either randomizing different utterances or by controlling these effects in a balanced design, depending on the overall objective of the experiment. Prerecorded test materials are typically used in order to control for these effects. The use of digitized speech materials has greatly increased the degree of control over these sources of variability. Nonsense syllables, however, provide no information on the prosodic characteristics of speech or normal discourse. As such, they cannot be used as measures of overall performance although, because of the high correlation between different aspects of speech perception, they can provide a rough estimate of overall communication performance (Fletcher, 1929).

Single-word tests are a step closer to, but still some distance from, everyday speech. Phonetically balanced word lists have been developed in order to provide test materials that are representative of the language at the word level (Egan, 1948; Hirsh et al., 1952). These lists can be used for both diagnostic and performance measures although the latter application is subject to several assumptions that are not always valid. Empirical evaluations show that for normal-hearing persons listening to speech in noise, a high correlation exists between test scores on well-designed phonetically balanced word lists and overall ability to understand speech (Egan, 1948). This correlation does not necessarily hold for all hearing-impaired persons or for persons listening to degraded speech using special-purpose sensory aids.

The use of meaningful words imposes some constraints on possible test formats. The very precise, closed-response-set format can be used but with difficulty in that it is not always possible to find meaningful words that meet all the design requirements of a well-balanced response set. The open-ended response format (ANSI-1960) is free of such constraints but subject to other limitations. Significant learning effects can occur on repeated measurements using the same word lists. Repeated measurements can be obtained on separate word lists of equal difficulty, but it is extremely difficult to develop matched sets of word lists. Word lists of equal difficulty under one set of conditions are not necessarily equally difficult under other conditions. Waltzman and Levitt (1978), for example, found significant interlist differences for the PAL PB-50 word lists when visual lipreading cues were available, but not for the audition-only condition. Similarly, word lists that have been equated in difficulty for normal-hearing listeners are not necessarily equally difficult for hearing-impaired persons.

The process of equating word lists for the hearing-impaired needs to take into account the fact that different types of hearing impairment affect speech cues in different ways. As a consequence, word lists may differ in relative difficulty between different types of impairment. This appears to be a particularly troublesome problem that has not been addressed properly. One of the great advantages of closed-response set testing is that the same test items can be used repeatedly (in random order) with only minor learning effects (Dubno & Dirks, 1982; Dubno, Dirks, & Langhofer, 1982; House et al., 1965). Given repeated mesurements of this type, it is also a relatively simple matter to analyze the error patterns obtained. An inherent limitation of such an analysis is that the observed error pattern is circumscribed by the set of allowable response alternatives.

Sentence length material provides a much closer approximation to everyday speech than balanced word lists. Both the phonetic and linguistic constraints of the language as well as contextual factors can be taken into account. The development of sentence lists for speech recognition testing is, however, a very difficult task, largely because so many additional variables need to be taken into account. A second problem involves the measurement of speech level. Different measures will be obtained depending on the type of instrument used (VU meter, level recorder, statistical averager), the response time of the instrument, and the criterion used to decide whether or not speech is present. Further, the slope of the performance-intensity function is usually very steep and as a consequence, a small change in signal level (e.g., 1 dB) can produce a substantial change in test score (e.g., 5–10 percentage points). The problem of uncertainty in the specification of speech level is thus compounded by the great sensitivity of the test material to inaccuracies in the measurement of level. A related problem is that the choice of stimulus levels is much more critical in the design of experiments involving very steep performance-intensity functions. A wrong choice will result in measured test scores being either close to zero or 100%, thereby providing very little new information. Adaptive procedures can be used to circumvent this problem, but particular attention must be paid to the choice of initial value and the control of step size. The earlier discussion on these issues is especially pertinent here.

A third methodological problem is that if two or more test words occur within a sentence, these test items are not statistically independent. Misperception of one test word in a sentence is likely to affect the perception of other test words in the same sentence. The effect of correlations between successive test items is to increase test-retest variability in a manner similar to that of reducing the number of independent items in a test.

A major practical problem in the use of conventional sentence material such as the PAL phonetically balanced sentence lists (Egan, 1948) or the CID/CHABA sentences of everyday speech (Davis & Silverman, 1970) is that learning effects are substantial and the same sentences should not typically be used more than once with any one subject. This presents serious logistical problems for repeated or routine testing using sentences, because the preparation of large numbers of matched sentence lists is extremely difficult. It is possible to use a closed-response-set format with sentences under limited conditions, as in diagnostic testing for the perception of prosodic features (Levitt & Resnick, 1978) or in identifying target sentences as in the Synthetic Speech Identification Test (Speaks & Jerger, 1965). Research currently in progress at CUNY (Levitt, 1982) is investigating the use of a forced-choice, true-false response format in measuring sentence perception. The binary form of the subject's response allows the technique to be automated, using an adaptive transformed up-down strategy that converges on a prescribed level of performance. The technique has been demonstrated to work well in measuring reading speeds of sentences using visual displays and is currently being evaluated as a tool in measuring speech recognition ability. The use of a closed-response set with sentences opens up the possibility of substantially learning effects on repeated measurements with the same test materials. The strategy being evaluated uses similar words in different contexts and different syntactic forms (declarative, passive, negative) so that the ability to memorize all of the test sentences in a list is of no advantage to the subject in deciding whether any given test item is true or false.

The use of continuous discourse provides an even closer approximation to everyday speech than single sentences, but the problems involved in the preparation and administration of this test material are even greter than that for sentences. The use of continuous discourse has been found to be quite practical in hearing aid selection using the paired-comparison procedure. In this case, the difficult problem of specifying level is one of the variables to be controlled. Basically, the technique consists of the subject listening to continuous discourse through two hearing aids and choosing the hearing-aid that, in the subject's opinion, is better acccording to a prescribed criterion (e.g., A is more intelligible than B); hence, many of the difficulties discussed above are circumvented. This technique has the very important practical advantage of great speed. A binary choice of this type is usually obtained in 10 s, as compared to the 5 or 10 min required for the administration of conventional speech recognition tests. The use of binary judgments also allows for the use of efficient adaptive strategies for converging on the best hearing aid.

The distinction between precision and accuracy is particularly important for this type of test. The test-retest repeatability of paired-comparison judgments between hearing aids is relatively good, exceeding that obtained with conventional speech recognition tests (Studebaker, 1982). At the same time, a few differences have been obtained between paired-comparison judgments of relative intelligibility and relative intelligibility as measured by conventional speech tests. There is no a priori reason for presuming that either type of measurement is more accurate, and it remains to be seen which method of measurement results in more satisfactory prescriptive fitting of hearing aids.

A technique which attempts to provide a performance measure that is more representative of actual interpersonal communication is the tracking method (DeFillipo & Scott, 1978). This technique, in essence, measures the time taken for a speaker to communicate a specified amount of information to a listener, such as reading a prescribed text. The listener repeats what is heard, requesting repetition or clarification of material that is misperceived. The cost of such interactions is an increase in the time taken for the task. This technique does not provide absolute measures of performance but rather relative measures using the same talker-listener pair. It is well suited for comparing communication channels or for evaluating the relative effect of a sensory aid in improving communication ability. The strength of the procedure lies in its providing a measure of performance that takes into account the interactive nature of human communication. There are, however, several methodological considerations that need further investigation. Firstly, the technique is sensitive to the choice of material. Talker-listener interactions may differ for different types of test material thereby obscuring the measurement of direct interest, the relative effect of the communication channel or sensory aid being evaluated. Occasionally, a single difficult word or phrase in an otherwise easy test paragraph may require many repetitions for clarification, resulting in a misleadingly low measure of overall performance. The difference between precision and accuracy is also more difficult to assess for problems of this type. These difficulties can be reduced if not eliminated by minor modifications to the procedure (e.g., developing standardized test material, setting an upper bound to the number of clarifications that can be requested), but the issues need to be addressed.

In summary, the basic principles underlying the measurement process have been discussed with particular reference to speech recognition testing. Diagnostic or analytic test procedures involving the segmental characteristics of speech appear to be well advanced and their implementation illustrates well the practical application of these principles. In contrast, the development of tests of relative performance is not as advanced. There are difficult methodological problems associated with the development of effective measures of performance, particularly when the test involves more complex material than single words or nonsense syllables. Methodological issues and choice of test material are closely intertwined. There is much to be done in the development of more effective methodologies for use with more realistic test material, such as sentences and continuous or interactive discourse.

REFERENCES

AMERICAN NATIONAL STANDARDS INSTITUTE. (1960). *American National Standard measurement for monosyllabic word intelligibility* (ANSI S3.2-1960). New York: ANSI.

DAVIS, H., & SILVERMAN, S. R. (Eds.). (1970). *Hearing and deafness* (3rd ed.). New York: Holt, Rinehart & Winston.

DEFILIPPO, C. L., & SCOTT, B. L. (1978). A method for training and evaluating the reception of ongoing speech. *Journal of the Acoustical Society of America, 63* 186–192.

DIXON, W. J., & MOOD, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association, 43,* 109–126.

DUBNO, J. R., & DIRKS, D. D. (1982). Evaluation of hearing-impaired listeners using a Nonsense Syllable Test. I. Test reliability. *Journal of Speech and Hearing Research, 25,* 135–141.

DUBNO, J. R., DIRKS, D. D., & LANGHOFER, L. R. (1982). Evaluation of hearing-impaired listeners using a Nonsense Syllable Test. II. Syllable recognition and consonant confusion patterns. *Journal of Speech and Hearing Research*, 25, 141–148.

EGAN, J. P. (1948). Articulation testing methods. *Laryngoscope, 58,* 955–991.

FLETCHER, H. (1929). *Speech and hearing* (1st ed.). New York: Van Nostrand.

HIRSH, I.J., DAVIS, H., SILVERMAN, S. R., REYNOLDS, E., ELDERT, E., & BENSON, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17,* 321–337.

HOUSE, A. S., WILLIAMS, C. E., HECKER, M. H. L., & KRYTER, K. D. (1965). Articulation testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America,* 37, 158–166.

KAMM, C. A., MORGAN, D. E., & DIRKS, D. D. (1983). Accuracy of adaptive procedure estimates of PB-max level. *Journal of Speech and Hearing Disorders, 48,* 202–209.

LEVITT, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467–477.

Levitt, H. (1978). Adaptive testing in audiology. *Scandinavian Audiology,* 6(Suppl.), 241-291.

Levitt, H. (1982). *Rehabilitation strategies for the hearing-impaired* (Annual Rep., PO1 NS 17764-02). New York: City University of New York.

Levitt, H. (1984). The phoneme: One of life's little uncertainties. In L. Raphael, C. B. Raphael, & M. Valdovinos (Eds.), *Language and cognition: Essays in honor of Arthur J. Bronstein.* New York: Plenum.

Levitt, H., & Resnick, S. B. (1978). Speech reception by the hearing-impaired: Methods of testing and the development of new tests. In C. V. Ludvigsen & J. Barford (Eds.), Sensorineural hearing impairment and hearing aids. *Scandinavian Audiology,* 6(Suppl.). 105–130.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics,* 22, 400–407.

Speaks, C., & Jerger, J. (1965). Methods for measurement of speech identification. *Journal of Speech and Hearing Research, 8,* 185–194.

Studebaker, G. A. (1982). Hearing aid selection: An overview. In G. A. Studebaker & F. H. Bess (Eds.), *The Vanderbilt hearing-aid report* (Monographs in Contemporary Audiology). Upper Darby, PA: Associated Hearing Instruments.

Swets, J. A. (Ed.). (1964). *Signal detection and recognition by human observers.* New York: Wiley.

Waltzman, S., & Levitt, H. (1978). The SIL as a predictor of face-to-face communication. *Journal of the Acoustical Society of America, 63,* 581–590.

Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society, B25,* 1–48.

Wetherill, G. B., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology, 18,* 1–10.

Chapter 12

SPEECH RECOGNITION TESTING: NONCONVENTIONAL TESTING TECHNIQUES AND APPLICATION TO CHILDREN

STEFFI B. RESNICK

*John F. Kennedy Institute
for Handicapped Children, Baltimore, MD*

Review of clinical techniques for assessing the speech recognition abilities of young children reveals a variety of measures whose validity, reliability, and sensitivity must be considered questionable. The majority of the clinical tests of speech recognition for children younger than 8 years of age are closed-response, word identification tests which require the child to select, from among a set of pictorial items, representations of the monosyllabic words that serve as the stimulus items. Criticisms of the test procedures reflect dissatisfaction with both the test stimuli and response format.

The dissatisfaction with monosyllabic word tests for adults is evident in concerns regarding the reliability (Chial & Hayes, 1974; Shore, Bilger, & Hirsh, 1960) and sensitivity (Jerger, Malmquist, & Speaks, 1966) of the measures as well as the difficulties in establishing equivalence of various forms of the test (Owens & Schubert, 1977). When monosyllabic words are used for children and the response mode is picture pointing, additional problems arise as a consequence of (a) the constraints on the size of the stimulus pool imposed by the child's receptive vocabulary (Elliott & Katz, 1980b); (b) the limits on the stimulus pool occasioned by the necessity for pictorial representation of the stimulus items and response alternatives (Goldman, Fristoe, & Woodcock, 1974); and (c) the confounding of measures of phoneme identification with the linguistic experience of the child (Elliott et al., 1979; Sanderson-Leepa & Rintelmann, 1976).

Recently, Wilson and Antablin (1980, 1982) have suggested that for normal adults picture-pointing responses to a monosyllabic word test yield lower scores than do word-pointing responses and that "additional cognitive processes are required to transform a picture into a lexical unit" (1980, p. 231). The role of cognitive processes in determining the performance of young children on word-recognition/picture-recognition tests has not been examined. The appropriateness of the pictorial representation may be of particular significance in determining the performance of the young child.

Clinicians and investigators alike, responding to the additional constraints imposed by the limited attention, cooperation, and attention of young children, have adopted various strategies for testing. Included among these are (a) live-voice presentation of materials (Ross & Lerman, 1970; Siegenthaler & Haspiel, 1966); (b) reduction in the number of test items (Erber, 1980); and (c) limitation of the number of response alternatives (Siegenthaler & Haspiel, 1966). All these approaches are regarded as undesirable from the standpoint of reliability and sensitivity.

Recognition of the limitations of these approaches has prompted the utilization of testing approaches involving other than a picture-pointing response for the evaluation of children with speech and/or language problems (Locke, 1980) and for the investigation of developmental changes in speech sound recognition (Menary, Trehub, & McNutt, 1982).

Attempts to circumvent problems associated with the need for pictorial representation of response alternatives have led some investigators to employ a paired-comparison task in which the child is asked to make a "same–different" judgment (Graham & House, 1971). However, Schwartz and Goldman (1974) concluded that the "paired-comparison" approach was particularly unsuitable for use with young children because error scores were greater when test items were presented using the paired-comparison task than when the test items were presented in a carrier phrase and a picture-pointing response was required. The differences were most marked when the materials were presented in noise. Unfortunately, interpretation of Schwartz and Goldman's data is complicated by the response required for the paired-comparison task. As described by the authors,

> For the paired-comparison context, each response plate contained three pairs of pictures, and the child was required to point to the pair of pictures named by the speaker. (p. 28)

For the young child with a hearing impairment, the limitations of current techniques for assessing speech recognition may compromise his/her audiologic evaluation and management and the educational programming (ASHA, 1974). The need for some measure of sound recognition ability for educational and audiological management of the severely or profoundly hearing-impaired child has prompted development of several tests which assess the child's ability to make discriminations based on syllable length, stress pattern, or vowel cues (Cramer & Erber, 1974; Erber 1980); to discriminate among familiar environmental sounds (Finitzo-Hieber, Gerling, Matkin, & Cherow-Skalka, 1980); or to duplicate rhythm and intonation patterns (Koike & Asp, 1981). These tests do not meet the needs of the less severely hearing-impaired child.

Issues related to test sensitivity and reliability appear to have received little attention in the development of these measures, and the procedures are of questionable value for evaluation of the child's ability to process speech signals.

Efforts to develop clinical tests of speech recognition ability appropriate for children other than the severely or profoundly hearing impaired have resulted in a variety of monosyllabic word tests in which the items are reported to be in the receptive vocabulary of young children (Elliott & Katz, 1980a; Goldman, Fristoe, & Woodcock, 1974; Haskins 1949; Ross & Lerman, 1970; Siegenthaler & Haspiel, 1966). Despite such efforts, the performance of normal-hearing children with normal intelligence on monosyllabic word tests has been shown to improve as a function of age, both in quiet (Elliott & Katz, 1980b; Sanderson-Leepa & Rintelmann, 1976) and in noise (Mills, 1975; Schwartz & Goldman, 1974). Elliott and her colleagues (1979) observed age-related changes in performance on their "adaptive speech understanding test procedure" when monosyllabic words within the receptive vocabulary of 3-year-olds were presented in quiet to 5- to 10-year-old children. They observed larger age-related changes on their closed-response task than on their open-response and suggested the differences may have resulted, in part, from "an increased ability of the older subjects to utilize minimal amounts of acoustic information regarding the vowels in the response alternatives and to achieve semantic closure" (p. 20). These investigators reported no age-dependent changes in performance in noise; they attributed the absence of the developmental changes under noise conditions to the overall reduction in the cues for semantic closure available to the older subjects under quiet conditions.

Elliott and Katz (1980a) recently have developed a picture-identification clinical test for use with children 3 years of age and older, the NU-CHIPS (Northwestern University Children's Perception of Speech). The performance of normal-hearing children on that test clearly is age dependent. Elliott and Katz have reported that for hearing-impaired children, receptive vocabulary age (as assessed using the Peabody Picture Vocabulary Test—Dunn, 1981) was a significant "predictor" of performance on the NU-CHIPS, but chronological age was not. Considered in this context, their conclusion that the construct validity of the test is demonstrated by the relation between pure-tone sensitivity and test performance by hearing-impaired children is of some concern.

The monosyllabic word tests for children and for adults have been developed with an emphasis on phonemic balance, despite the limited sensitivity of phonemically balanced materials to conditions which might be expected to degrade recognition of the speech materials. In response to the limited sensitivity to audiometric configuration of those materials for adults, word lists weighted with items frequently misperceived by hearing-impaired listeners or by normal-hearing listeners under difficult listening conditions have been developed. The sensitivity of the high-frequency loaded word lists to changes in the frequency reponse of amplification systems has been demonstrated for listeners with high-frequency hearing loss (Pascoe, 1975; Skinner, 1980). The 50-item word lists employed by Pascoe and Skinner are presented in a closed-response format. The response alternatives consist of a 50-item pool from which the items themselves are drawn. The practice effects reported for those materials has precluded their clinical application.

The California Consonant Test (CCT) developed by Owens and Schubert (1977) for clinical administration to adults is similarly weighted with items of demonstrated difficulty for listeners with high-frequency hearing impairment. The CCT is a 100-item multiple-choice test of consonant identification. Each item has four response alternatives which differ only with respect to their final or initial consonant. Research on its sensitivity to systematic changes in listening conditions is limited (Schwartz & Surr, 1979), but development of the materials appears to have been pursued with careful attention to item selection and list construction. The University of Oklahoma Closed Response Speech Test, developed for clinical use with severely hearing-impaired children 9 years of age and older (Jones & Studebaker, 1974), has several features in common with the CCT, including the disregard of phonemic balance and the presentation of response foils which differ from the stimulus item primarily with respect to place of articulation.

If an increasing ability to achieve semantic closure is a major determinant of age-related changes in performance on word identification tasks, then word materials may be particularly inappropriate for assessing peripheral processing for children whose linguistic experience has been limited as a consequence of age, hearing loss, mental retardation, auditory processing disorder, a non-English language background, inadequate language stimulation, or some combinations of those factors. Therefore, the value of investigations employing word identification tests for the purposes of documenting reduced speech sound discrimination abilities in children with a history of middle-ear pathology (Brandes & Ehinger, 1981), with language impairment (Weiner, 1969), or with reading problems (Flynn & Byrne, 1970) must be viewed as questionable.

However, the use of materials that are linguistically loaded to a greater degree may be more appropriate, if the goal is to document the extent to which cognitive processing of speech is affected for various clinical populations. Recent work by Elliott (1979) provides evidence of the role of age in determining the intelligibility of linguistically loaded materials under difficult listening conditions. Administration of the Speech Perception in Noise Test (SPIN—Kalikow, Stevens, & Elliott, 1977) to 9- to 17-year-olds in quiet and under various speech-to-babble ratios revealed an age-by-sentence predictability interaction due to performance on the high-predictability sentences at the 0 speech-to-babble ratio. Performance of 11- and 13-year-olds was poorer than performance of 15- and 17-year-olds on the high-predictability sentences but not on the low-predictability sentences. Additional testing of normal 9-year-olds and 9-year-olds with "learning problems" showed that "9-year-olds with learning problems performed more poorly than the normally progressing 9-year-olds" (Elliott, 1979, p. 652). They also were found to have significantly lower scores on the Peabody Picture Vocabulary Test.

If, however, the goal is to document the peripheral processing of the speech signal, with minimal reliance on cognitive processing, the linguistically loaded materials may be inappropriate for children. If the cues for semantic closure and word frequency effects can be minimized through selection of alternative stimulus materials (e.g., nonsense syllables), then changes in test performance that are related to age or linguistic experience should be reduced. Units of speech that have a minimal linguistic load (i.e., nonsense syllables) have been employed routinely in the experimental investigation of the speech perception abilities of infants and young children (Abbs & Minifie, 1969). Such investigation has necessitated development of stimulus and response paradigms which are capable of reflecting the auditory discrimination capabilities of the subjects, yet are appropriate to the level of their sensory, perceptual, and motor development (Eimas, 1974); Kuhl, 1979; Morse, 1978). In addition, nonsense syllables (produced as natural utterances or computed-synthesized signals) have been applied increasingly to the study of developmental changes in speech perception in children (Bernstein, 1979).

Nonsense syllables have been used to assess the effects of manipulation of the conditions of signal presentation on patterns of phoneme identification errors in normal-hearing adult listeners (Miller & Nicely, 1955; Wang & Bilger, 1973). Recently, nonsense syllable materials have gained application in studies of the effects of both audiometric configuration (Bilger & Wang, 1976; Kamm, Dirks, & Carterette, 1982; Walden, Schwartz, Montgomery, & Prosek, 1981; Wang, Reed, & Bilger, 1978) and amplification on the phoneme identification errors of hearing-impaired adults. Nonsense syllable materials are useful in those studies because they obviate the need for construction of equivalent test forms. The items may be identical on all forms of the test; only the order of syllable presentation need be changed (Resnick et al., 1976). Presentation of a common pool of nonsense syllable items apparently does not require utilization of complex schemes for distribution of practice effects required, when repeated use is made of a common pool of monosyllabic words which have a high probability of being misperceived (Pascoe, 1975; Skinner, 1980).

The City University of New York Nonsense Syllable Test (CUNY NST), developed by Resnick, Dubno, Hoffnung, and Levitt (1975), is gaining acceptance in research applications with adult normal-hearing and hearing-impaired listeners as a consequence of the reliability of the test results (Dubno & Dirks, 1982) and of the sensitivity of the measures to systematic manipulation of the conditions of signal presentation (Dubno & Levitt, 1981; Levitt, Collins, Dubno, Resnick, & White, 1978). The nonsense syllables comprising the CUNY NST were selected to include all of the major consonantal sounds in combination with one or more of the three vowels occupying extreme positions on the vowel triangle /i/, /ɑ/ and /u/. The syllables are organized into sets of seven to nine syllables each. Within each set the following are maintained constant: (a) the class of consonant represented (i.e. voiced or voiceless); (b) the position of the consonant within the syllable; and (c) the vowel context. Each set is administered as a unit. The test is a closed-response test with the response alternatives for any item within a set corresponding to all the syllables within that subtest. The subtests were constructed to include the most frequent confusions reported for hearing-impaired listeners with mild-to-moderate hearing loss (Bilger & Wang, 1976; Owens, Benedict, & Schubert, 1972) and for normal-hearing listeners under difficult listening conditions (Miller & Nicely, 1955; Wang & Bilger, 1973).

Efforts to establish the reliability of the NST appear somewhat misdirected. Dubno and Dirks (1982) demonstrated that repeated administration of the test under identical conditions (speech at 90 dB SPL, cafeteria noise at 70 dB SPL) yielded repeatable performance by listeners with sensorineural hearing-impairment. Dubno and her colleagues (Dubno, Dirks, & Langhofer, 1982) suggested, following their analysis of the errors of hearing-impaired listeners on the Nonsense Syllable Test (Levitt et al., 1978), that the consistency of the responses would be "especially important . . . if the purpose of the evaluation is to monitor changes in consonant recognition resulting from spectral shaping by an amplification system" (Dubno et al., 1982, p. 148). Unfortunately, their work and that of other investigators have failed to document the utility of the CUNY NST for that purpose. That is, the sensitivity of the measures to systematic manipulation of the spectrum and the repeatability of that sensitivity have not yet been demonstrated. Nor, perhaps more importantly, have the productions of these nonsense syllables been shown to be representative of those the hearing-impaired listener may encounter. To the extent that the error patterns reflect the characteristics of the particular talker and procedures for recording and presentation, as preliminary evidence suggests (Levitt et al., 1978), the value of the particular test must be questioned.

Nonsense syllable materials do appear to offer several advantages for assessing phoneme identification errors in children. Bess and Gibler (1981) have reported success in applying the NST to children between 6 and 13 years of age. The development of a nonsense syllable test for young children will require additional investigation of the error patterns of children. Review of the literature suggests that there are not sufficient data to support an assumption that children and adults make similar errors in phoneme identification. Graham and House (1971) concluded, on the basis of their experimental results for girls from 3 to 4½ years of age who provided "same-different" responses to consonant pairs, that "the perceptual behavior of the children is similar to that of adults, except that the children produced more errors than an adult is expected to make in a comparable task" (p. 565). The consonants were embedded in /hə'Cɑd/ɚ and were presented live-voice. Bernstein (1982), however, suggested that comparison of the data of Graham and House with the data acquired by Miller and Nicely (1955) for adults on a labeling or identification procedure revealed differences in the types of errors made by listeners in the two studies. Wang and Bilger (1973), who reported patterns of confusion for adults on a labeling task similar to those reported by Miller and Nicely, suggested that the "rather atypical" results of Graham and House may reflect either the subject population or the techniques applied to reanalyses of the "same-different" data.

Additional support for differences in speech sound identifiction between young children and adults derives from developmental studies of speech perception (Elliott, Longinotti, Meyer, Raz, & Zucker, 1981; Krause, 1978; Zlatin & Koenigsnecht, 1975) which provide evidence of changes on the position and steepness of categorical boundaries with age. Graham and House (1981) suggested that the children in their study "may have been responding to small but perceptible differences which would have been ignored had their phonological systems been further developed" (p. 563).

Considered in the context of investigations of speech perception in young children, selection of appropriate stimulus items for examination of the performance of young children on a nonsense syllable task becomes critical, as do decisions regarding (a) use of a carrier phrase, (b) procedures for level equalization, and (c) techniques for stimulus presentation and response elicitation.

REFERENCES

Abbs, M. S., & Minifie, F. D. (1969). Effect of acoustic cues on fricatives on perceptual confusions in preschool children. *Journal of the Acoustical Society of America, 46,* 1535–1542.

American Speech and Hearing Association, Committee on Rehabilitative Audiology. (1974). The audiologist: Responsibilities in the habilitation of the auditorily handicapped. *Asha, 16,* 14–18.

Bernstein, L. E. (1979). Developmental differences in labeling VOT continua with varied fundamental frequency. *Journal of the Acoustical Society of America, 65*(Suppl. 1), S1.

Bernstein, L. E. (1982). Ontogenetic changes in children's speech-sound perception. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice.* New York: Academic Press.

Bess, F. H., & Gibler, A. M. (1981). Syllable recognition skills of unilaterally hearing-impaired children. *Asha, 23,* 724.

Bilger, R. C., & Wang, M. D. (1976). Consonant confusions in patients with sensorineural hearing loss. *Journal of Speech and Hearing Research, 19,* 718–748.

Brandes, P. J., & Ehinger, D. M. (1981). The effects of early middle ear pathology on auditory perception and academic achievement. *Journal of Speech and Hearing Disorders, 46,* 301–307.

Chial, M. R., & Hayes, C. S. (1974). Hearing aid evaluation methods: Some underlying assumptions. *Journal of Speech and Hearing Disorders, 39,* 270–279.

Cramer, K. D., & Erber, N. P. (1974). A spondee recognition test for young hearing-impaired children. *Journal of Speech and Hearing Disorders, 39,* 304–311.

Dubno, J. R., & Dirks, D. D. (1982). Evaluation of hearing-impaired listeners using a Nonsense Syllable Test. I. Test reliability. *Journal of Speech and Hearing Research, 25,* 135–141.

Dubno, J. R., Dirks, D. D., & Langhofer, L. R. (1982). Evaluation of hearing-impaired listeners using a Nonsense Syllable Test. II. Syllable recognition and consonant confusion patterns. *Journal of Speech and Hearing Research, 25,* 141–148.

Dubno, J. R., & Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *Journal of the Acoustical Society of America, 69,* 249–261.

Dunn, L. M. (1981). *Peabody Picture Vocabulary Test.* Circle Pines, MN: American Guidance Service.

Eimas, P. D. (1974). Linguistic processing of speech by young infants. In R. L. Schiefelbusch & L. L. Lloyd (Eds.). *Language perspectives acquisition, retardation and intervention.* Baltimore: University Park Press.

Elliott, L. L. (1979). Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability. *Journal of the Acoustical Society of America, 66,* 651–653.

Elliott, L. L., & Katz, D. R. (1980a). *Northwestern University Children's Perception of Speech (NU-CHIPS): Technical manual.* St. Louis: Auditec.

Elliott, L. L., & Katz, D. R. (1980b). *Development of a new children's test of speech discrimination.* St. Louis: Auditec.

Elliott, L. L., Connors, S., Kille, E., Levin, S., Ball, K., & Katz, D. (1979). Children's understanding of monosyllabic nouns in quiet and noise. *Journal of the Acoustical Society of America, 66,* 12–21.

Elliott, L. L., Longinotti, C., Meyer, D., Raz, I., & Zucker, K. (1981). Developmental differences in identifying and discriminating CV syllables. *Journal of the Acoustical Society of America, 70,* 669–677.

Erber, N. P. (1980). Use of the auditory numbers tests to evaluate speech perception abilities of hearing-impaired children. *Journal of Speech and Hearing Disorders, 45,* 527–532.

Finitzo-Hieber, T., Gerling, I. J., Matkin, N. D., & Cherow-Skalka, E. (1980). A sound effects recognition test for the pediatric audiological evaluation. *Ear and Hearing, 1,* 271–276.

Flynn, P. T., & Byrne, M. C. (1970). Relationship between reading and selected auditory abilities of third-grade children. *Journal of Speech and Hearing Research, 13,* 725–730.

Goldman, R., Fristoe, M., & Woodcock, R. (1974). *The Goldman-Fristoe-Woodcock Auditory Skills Test Battery.* Circle Pines, MN: American Guidance Service.

Graham, L. W., & House, A. S. (1971). Phonological oppositions in children: A perceptual study. *Journal of the Acoustical Society of America, 49,* 559–566.

Haskins, H. A. (1949). *A phonetically balanced test of speech discrimination for children.* Unpublished master's thesis, Johns Hopkins University, Baltimore.

Jerger, J., Malmquist, C., & Speaks, C. (1966). Comparison of some speech intelligibility tests in the evaluation of hearing aid performance. *Journal of Speech and Hearing Research, 9,* 253–258.

Jones, K. O., & Studebaker, G. A. (1974). Performance of severely hearing-impaired children on a closed-response, auditory speech discrimination test. *Journal of Speech and Hearing Research, 17,* 531–540.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61,* 1337–1351.

Kamm, C. A., Dirks, D. D., & Carterette, E. C. (1982). Some effects of spectral shaping on recognition of speech by hearing-impaired listeners. *Journal of the Acoustical Society of America, 71,* 1211–1224.

Koike, J. M., & Asp, C. W. (1981). Tennessee Test of Rhythm and Intonation Patterns. *Journal of Speech and Hearing Disorders, 46,* 81–87.

Krause, S. E. (1978). *Developmental use of vowel duration as a cue to postvocalic consonant voicing: A perception and production study.* Unpublished doctoral dissertation, Northwestern University, Chicago.

Kuhl, P. (1979). The perception of speech in early infancy. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice.* New York: Academic Press.

Levitt, H. Collins, M. J., Dubno, J. R., Resnick, S. B., & White, R. E. C. (1978). *Development of a protocol for the prescriptive fitting of a wearable master hearing aid* (CUNY Research Rep. 11). New York: Communication Science Laboratory.

Locke, J. L. (1980). The inference of speech perception in the phonologically disordered child. Part II: Some clinically novel procedures, their use, some findings. *Journal of Speech and Hearing Disorders, 45,* 445–468.

Menary, S., Trehub, S. E., & McNutt, J. (1982). Speech discrimination in preschool children: A comparison of two tasks. *Journal of Speech and Hearing Research, 25,* 202–207.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Mills, J. H. (1975). Noise and children: A review of the literature. *Journal of the Acoustical Society of America, 58,* 767–779.

Morse, P. A. (1978). Infant speech perception: Origins, processes, and Alpha Centauri. In F. D. Minifie & L. L. Lloyd (Eds.), *Communicative and cognitive abilities—Early behavioral assessment.* Baltimore: University Park Press.

Owens, E., Benedict, M., & Schubert, E. D. (1972). Consonant phonemic errors associated with pure-tone configurations and certain kinds of hearing-impairments. *Journal of Speech and Hearing Research, 15,* 308–322.

Owens, E., & Schubert, E. D. (1977). Development of the California Consonant Test. *Journal of Speech and Hearing Research, 20,* 463–474.

Pascoe, D. P. (1975). Frequency responses of hearing aids and their effects on the speech perception of hearing-impaired subjects. *Annals of Otology, Rhinology, and Laryngology, 84*(Suppl. 23).

RESNICK, S. B., DUBNO, J. R., HOFFNUNG, S., & LEVITT, H. (1975). Phoneme errors on a nonsense syllable test. *Journal of the Acoustical Society of America, 58*(Suppl. 1), S114.

RESNICK, S. B., DUBNO, J. R., HAWIE, D. G., HOFFNUNG, S., FREEMAN, L., & SLOSBERG, R. M. (1976). *Phoneme identification on a closed response nonsense syllable test.* Paper presented at the Annual Convention of the American Speech and Hearing Association, Houston.

ROSS, M., & LERMAN, J. W. (1970). A picture identification test for hearing-impaired children. *Journal of Speech and Hearing Research, 13,* 44–53.

SANDERSON-LEEPA, M. E., & RINTELMANN, W. F. (1976). Articulation functions and test-retest performance of normal-hearing children on three speech discrimination tests: WIPI, PBK-50, and NU Auditory Test No. 6. *Journal of Speech and Hearing Disorders, 41,* 503–519.

SCHWARTZ, A. H., & GOLDMAN, R. (1974). Variables influencing performance on speech-sound discrimination tests. *Journal of Speech and Hearing Research, 17,* 25–32.

SCHWARTZ, D. M., & SURR, R. K. (1979). Three experiments on the California Consonant Test. *Journal of Speech and Hearing Disorders, 44,* 61–72.

SHORE, L., BILGER, R., & HIRSH, I. (1960). Hearing aid evaluation: Reliability of repeated measurements. *Journal of Speech and Hearing Disorders, 25,* 152–167.

SIEGENTHALER, B., & HASPIEL, G. (1966). *Development of two standardized measures of hearing for speech by children* (Project No. 2372, Contract No. OE-5-10-003). Washington, DC: U.S. Department of Health, Education and Welfare.

SKINNER, M. W. (1980). Speech intelligibility in noise-induced hearing loss: Effects of high-frequency compensation. *Journal of the Acoustical Society of America, 67,* 306–317.

WALDEN, B. E., SCHWARTZ, D. M., MONTGOMERY, A. A., & PROSEK, R. A. (1981). A comparison of the effects of hearing impairment and acoustic filtering on consonant recognition. *Journal of Speech and Hearing Research, 24,* 32–43.

WANG, M. D., & BILGER, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America, 54,* 1248–1266.

WANG, M. D., REED, C. M., & BILGER, R. C. (1978). A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusion. *Journal of Speech and Hearing Research, 21,* 5–36.

WEINER, P. S. (1969). The cognitive functioning of language deficient children. *Journal of Speech and Hearing Research, 12,* 53–64.

WILSON, R. H., & ANTABLIN, J. K. (1980). A picture identification task as an estimate of the word-recognition performance of nonverbal adults. *Journal of Speech and Hearing Disorders, 45,* 223–237.

WILSON, R. H., & ANTABLIN, J. K. (1982). The Picture Identification Task, A reply to Dillon. *Journal of Speech and Hearing Disorders, 47,* 111–112.

ZLATIN, M. A., & KOENIGSKNECHT, R. A. (1975). Development of the voicing contrast: Perception of stop consonants. *Journal of Speech and Hearing Research, 18,* 541–553.

Chapter 13

MEASUREMENT OF CENTRAL AUDITORY INTEGRITY

CHARLES E. SPEAKS

University of Minnesota, Minneapolis

Integrity and function of the central auditory system have been investigated for more than 20 years by the technique of dichotic stimulation with either speech or other complex signals. Perhaps these two decades of research and the voluminous literature that has been written may best be described as chaotic and, in some cases, unscientific. There is, for example, nothing approaching universal agreement on fundamental issues that contribute to acceptable reliability of the procedures. By way of sharp contrast, there appears to be nearly universal acceptance that the many different procedures that are used lead to a valid assessment of central auditory function.

Discussion about this area of speech recognition has been organized into two rather broadly framed problems: (a) use of dichotic speech tests to assess hemispheric dominance (specialization) in listeners with normal brain function; and (b) clinical use of dichotic speech tests to assess central auditory deficit.

ASSESSMENT OF HEMISPHERIC DOMINANCE WITH DICHOTIC TESTS

Justification for the use of dichotic tests to assess characteristics of normal central auditory function is predicated mainly on a single unverified assumption: When listeners with apparently normal brain function are tested dichotically, the ear with the higher score is assumed to be contralateral to the hemisphere that is dominant for processing the class of signals presented. This translates to an expectation of a right-ear advantage (REA) for speech signals to reflect the left-brain dominance that apparently exists for more than 90% of the population, and to an expectation of a left-ear advantage (LEA) when signals are used that require processing mechanisms of the right hemisphere. Some investigators have pushed the assumption even further by asserting that the size of the observed ear advantage reflects the degree to which brian dominance has been established. There appear to be several obvious problems in this area that have not been addressed adequately and are described rather briefly.

Size of Observed Ear Advantage

Most investigators report a mean right-ear advantage for speech and that mean REA almost always lies somewhere between 4–6% at the low end and 8–12% at the high end. In research conducted at Minneapolis, the mean REA for a group of listeners is about 6%, and for two thirds of the listeners the advantage will lie between an LEA of 7% and an REA of 20%. Thus, an ear/brain hypothesis is not supported by a very large ear advantage; the presumed strong brain dominance produces only a very small advantage of one ear over the other. Furthermore, advantages as large as the mean of 6% only occur for signals such as CV nonsense syllables, signals that result in a fairly difficult signal-recognition task for the listener. If digits or words are used, there typically is no ear advantage because scores for both ears usually approach 100% correct.

Reversals in Direction of Ear Advantage

Several researchers have observed that the direction of ear advantage, right or left, observed from one listening run may be reversed on a second run (Blumstein, Goodglass, & Tartter, 1975; Pizzamiglio, DePascalis, & Vignati, 1974). That circumstance is difficult to interpret within the context of a strict ear/brain hypothesis. This problem was explored in an experiment (Speaks, Niccum, & Carney, 1982) in which 20 estimates of the ear advantage were obtained from each of 24 listeners; each estimate was based on 30 pairs of syllables. Seventeen of the 24 listeners reversed the direction of ear advantage from their own mean on one or more of the 20 listening runs.

These reversals do not appear in any way related to ear/brain relations. They arise from a small mean/sigma ratio for an ear advantage that is distributed normally for the individual listener. Figure 1 illustrates the situation by reference to three hypothetical listeners for whom multiple estimates of the ear advantage are made (Speaks et al., 1982). The mean advantages vary considerably among listeners. There are two properties of the advantage, however, that characterize virtually every listener who was tested: (a) the run-to-run estimates of the advantage are distributed normally; and (b) intra-listener variability in observed advantage is nearly the same for all listeners ($\sigma = 11\%$). From these two observations it stands to reason that the two listeners in Figure 1 with large ear advantages will rarely exhibit reversals because the $\bar{x}/\sigma$ ratio approaches 4:1. The listener in the middle with a rather typical ear advantage of 6% will evidence numerous reversals because the ratio is on the order of 0.5:1. From two experiments (Speaks & Niccum, 1977; Speaks et al., 1982), listeners

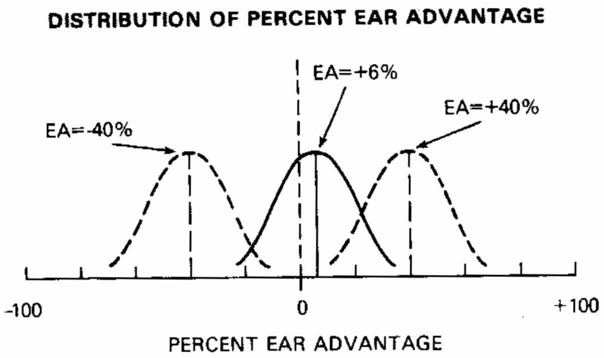

FIGURE 1. Illustration of three hypothetical listeners for whom multiple estimates of the ear advantage are made.

who do not show reversals have ratios on the order of 1.75:1 or so, whereas when the ratio is more like 0.3:1 reversals are common.

Differences Among Listeners

Generally, the proportion of the population observed to have a right-ear advantage should approximate the proportion of the population known to be left-brain dominant for speech if a strict ear/brain hypothesis is to be accepted. In the experiment (Speaks et al., 1982) already referred to, however, 18 of the 24 listeners had a mean observed REA, but that advantage was significant for only 12 of the listeners; there was a significant LEA for three of the listeners, and for the remaining nine listeners the difference between the left- and right-ear scores was not significant.

It is difficult to compare these results with those observed by other investigators because tests of significance for the individual listener are rarely reported. Although the Minneapolis group interprets the failure to be inconsistent with a strict ear/brain hypothesis, Wexler, Halwes, and Heninger (1981) would apparently disagree. On the surface, their outcome was similar; 31 listeners were tested and 14 had an REA, one had an LEA, and the advantage was not significant for the remaining 16 listeners. Their interpretation of this circumstance, however, was quite different. They argued that because 14 of the 15 listeners who had a significant ear advantage had a significant right-ear advantage, the findings were compatible with the reported incidence of language dominance. In other words, they seem to assume that processing must always be lateralized and that it is measurement error that results in a failure to observe an ear advantage. In fact, they postulated that decreasing measurement error by increasing the length of the test would lead to significant ear advantages for a larger proportion of the listeners tested. That was not found that to be the case (Speaks et al., 1982) even when the length was increased to 20 runs of 30 syllables per run; even in that situation, only about half of the listeners were expected to have a significant right-ear advantage. For this reason, the Minneapolis group agrees with Lauter (1982) that it is premature to speculate about the neurologic bases of responses to dichotic stimulation.

Reliability of Measurement

For the most part, reports of the reliability with which the ear advantage is assessed have been unsatisfactory: .80 by Ryan and McNeil (1974) for a 60-trial test; .74 by Blumstein et al. (1975) for an 80-trial test; .64 by Catlin, VanDerveer, and Teicher (1976) for a 120-trial test; and .66 by Speaks and Niccum (1977) for a 60-trial test. It may be that the common practice of administering only one or two listening runs of 30 pairs of signals per run will almost inevitably produce an unreliable estimate of the ear advantage. In the experiment by Speaks et al. (1982), with one listening run the split-half reliability is .62 and the standard error is about 13%, but with six listening runs of 30 pairs per run, the split-half reliability is .91 and the standard error is about 5.7%. It should be emphasized, however, that even with as many as six or more listening runs, achievement of a satisfactory level of reliability will not prevent reversals in direction of the observed ear advantage from run to run for the individual listener. If the mean/sigma ratio is small, reversals will occur.

Nature of the Listener's Task

The following is a description of the conventional testing paradigm that seems to be most popular as a two-ear monitoring, two-response procedure. The listener receives a pair of signals on a given listening trial, is instructed to attend equally to both ears, and is required to select two responses from a known set of alternatives—usually six. It may or may not be important to note that the listener is not required to assign either response to the ear in which he/she thinks the signal was heard. In some laboratories (Repp, 1977) the listener is asked to attend to both ears, but to provide only one response. Although the one-response procedure may have some methodological advantages, the following concerns apply equally to the two-response and one-response methods.

With either form of the conventional paradigm, the experimenter has no control over or knowledge of factors such as attentional bias, decision criteria, or other proclivities of the listener, and the experimenter certainly has no knowledge of the extent to which these listener variables contribute to both the direction and size of the observed ear advantage. Two experiments (Hayden, Kirstein, & Singh, 1979; Speaks & Ciccarelli, 1975) have shown that listeners can attend selectively to one or the other of the two ears. If the listener is asked to attend only to the right ear, one obtains a high score for responses that correspond to syllables presented to the right ear and a low score for syllables that had been presented to the left ear—There is a large REA. Alternatively, if the listener is asked to attend to the left ear, a large LEA is the result. Thus, with the conventional two-ear monitoring procedure where the listener is at least capable of attending more to one ear than to the other rather than follow the instructions, the experimenters do not know if the subjects did in fact adopt such a strategy, and of course they do not know the extent to which the observed ear advantage reflects a true difference in sensory capacity to the exclusion of such attentional biases.

For the past couple of years, the Minneapolis group has tried several variations in the testing procedures that reflect attempts to neutralize the effects of listener proclivities as much as possible. It may be an important by-product that with the procedure currently in use, the exact same paradigm may be used for both speech and nonspeech signals. Known as a target-monitoring paradigm (TARMON), the essential characteristics of the method are shown in Table 1. On a given listening run, a single syllable (e.g. /pɑ/ is designated as the target syllable and the other five syllables are the maskers. (It turns out that the term *masker* is probably inappropriate, but an acceptable alternative has not yet been discovered.) The listener knows the identity of the target, and the task is to listen for the target in each trial of the run and to engage in a Yes/No task. The listener responds "yes" to vote that the target was present and "no" that the target was not present. In the author's experiments, the a priori probability of target present is 0.5, and 40 trials per listening run are required to achieve a complete balancing of target with each of the five maskers. A minimum experiment requires 12 listening runs so that each of the six syllables serves as target for each of the two ears, although typically, a 24-run experiment is used. In a recently completed experiment (Katuski, Speaks, Penner, & Bilger, 1984), 20 listeners were tested with 24 runs. The results are expressed in a 2 × 2 matrix for each ear; the scores for each ear are given by $P(C)$ max, a d'-based statistic; the ear advantage is expressed by $P(C)$ max$_{\mathrm{RE}}$ - $P(C)$ max$_{\mathrm{LE}}$; and, in addition, listener criterion is described by β.

TABLE 1. Target-monitoring paradigm for a Yes/No task.

One Listening Run

Signal: /p/
Signal Present in LE: 20 trials
Signal Absent: 20 trials
40 Trials arranged in random order

Signal Present trials			Signal Absent trials		
Trial	LE	RE	Trial	LE	RE
1	p	t	21	t	k
2	p	t	22	t	b
3	p	t	23	t	d
4	p	t	24	t	g
5	p	k	25	k	t
6	p	k	26	k	b
7	p	k	27	k	d
8	p	k	28	k	g
9	p	b	29	b	t
10	p	b	30	b	k
11	p	b	31	b	d
12	p	b	32	b	g
13	p	d	33	d	t
14	p	d	34	d	k
15	p	d	35	d	b
16	p	d	36	d	g
17	p	g	37	g	t
18	p	g	38	g	k
19	p	g	39	g	b
20	p	g	40	g	d

Minimum experiment

(6 Targets × 2 Ears = 12 Runs)

Outcome

		Response	
		Yes	No
Signal	Present	□	□
	Absent	□	□

The mean ear advantage for the group was 6.1% with a standard deviation among listeners of 8.0%. A significant REA was observed for only 11 of the 20 listeners, and a significant LEA was obtained for 2 of the listeners. Although this mean of 6.1% was virtually the same as the value observed in previous experiments with the conventional procedure, it may be important to note that the interlistener standard deviation of 8.0% is much smaller. In other words, the occasional large REAs and LEAs that do occur with the conventional two-ear monitoring, two-response method were not observed.

Because the listeners in this experiment were tested only with the target-monitoring task, obviously one cannot specify how these advantages given by $P(C)$ max would compare with those obtained with the more traditional procedure in which listener criterion is not controlled. One could speculate about the relation, however, by comparing the ear advantage obtained from $P(C)$ max with that given by hit rate. With hit rate, the mean advantage is 12.3% and the standard deviation among listeners is 16.9%. In other words, the more criterion-free measure of ear advantage apparently reduced the size of the observed ear advantage by a factor of approximately 2. To the extent that our current methods are acceptable, the outcome is not compatible with the ear/brain hypothesis. Only 55% of the listeners had a significant REA when that REA is calculated over all six syllables serving as target. Furthermore, if we ask about the ear advantage for individual target syllables that is calculated over all 20 listeners, the advantage was significant for only two of the six syllables. If these failures can be attributed to measurement error, the nature of that error has not been discovered.

CLINICAL USE OF DICHOTIC SPEECH TESTS

There seem to be a few problems that exist with clinical application of dichotic speech tests. For the most part, clinical application of such tests has been based on the principle that the score for messages presented to the ear contralateral to the temporal lobe lesion will usually be substantially lower than the score for messages presented to the ear ipsilateral to the lesion. In the interest of brevity, the so-called *paradoxical ipsilateral-ear effect* that most certainly complicates clinical interpretation of the outcome of testing is not discussed (Sparks & Geschwind, 1968). This principle may be thought of as the *lesion effect* in the sense that the lower score for the contralateral ear is thought to be related to the side of the lesion (Damasio & Damasio, 1979; Linebaugh, 1978; Niccum, Rubens, & Speaks, 1981; Schulhoff & Goodglass, 1969; Sparks, Goodglass, & Nickel, 1970). There is another school of thought, however, in which the differences in scores for the two ears for the patient are interpreted similarly to the ear advantages observed for listeners who do not evidence brain damage. This is usually referred to as the *dominance effect* rather than a *lesion effect* (Johnson, Sommers, & Weidner, 1977; Moore & Weidner, 1975; Pettit & Noll, 1979). Thus, even though central auditory deficits exist, according to the dominance theory a low score for the right ear is described as a left-ear advantage and is thought to reflect the fact that the right hemisphere has assumed a dominant processing role consequent to the damage to the left hemisphere. According to the lesion theory, the same outcome is termed a right-ear deficit rather than a left-ear advantage and is thought to result from the damage located within the left hemisphere. No experiments are reported that have provided a satisfactory mechanism for sorting between these two opposing hypotheses, and one could partially subscribe to the posture expressed by Schulhoff and Goodglass (1969):

> Ear asymmetry under dichotic conditions is an index to lateral dominance in normal subjects. However, after unilateral brain injury, the "lesion effect" may interact with and possibly override the premorbid ear asymmetry so that dominance can no longer be inferred. (p. 157)

There are several factors that potentially may confound interpretation of clinical dichotic speech tests, and among those are the ipsilateral ear effect, size of the ear asymmetry, presence of aphasia, and coexistence of central auditory deficits and peripheral hearing loss (Speaks, 1980). Because the focus of this report is on speech recognition by the hearing-impaired, two problem areas should be addressed which are important when the patient for whom a central auditory test is contemplated also has a peripheral hearing loss. Both areas are tied to the fundamental question of the extent to which the observed asymmetry in scores for the two ears reflects the central deficit to the exclusion of the peripheral lesion.

Choice of Speech Test

Many different dichotic speech tests have been described in the literature, and among the more common are CV nonsense syllables, digits, familiar nouns, and sentences. Choice of test seems to reflect investigator preference more than an understanding of which test is likely to be most sensitive to the presence of central deficit but least likely to be influenced by the existence of a peripheral hearing loss. Attempts have been made to address this problem in two experiments. In one (Niccum et al., 1981) the patterns of performance obtained from 16 aphasic listeners on five verbal dichotic listening tests were compared: the highly overlearned digits, the more abstract CV nonsense syllables, and three word tests that comprised highly familiar, concrete nouns. One word test was termed *high contrast* because the six words differed from each other in terms of both consonantal and vocalic information. Another was a *vowel-word test* that required recognition of vocalic segments, and the third, a *consonant-word test*, required recognition of consonantal information. The signals in each of the five tests are shown in Table 2. Differences among results for the digits, high-contrast, and vowel-word tests were not significant. For the listeners, mean right-ear deficits on the order of 38–44% were observed and performance level ranged from 74 to 78%. The size of the ear asymmetry was about the same for the consonant-word test, but the task was more difficult; performance level dropped to about 65%. As expected, performance level was even lower for the CV syllables (50%), and a smaller mean right-ear deficit of 22% was observed.

On the surface, these results do not provide a certain basis for selection of one test over another. From a different perspective, however, the digits did emerge as potentially more desirable because of the relation of the digit scores to evidence of lesion location. The distribution of right-ear (RE) scores among patients was discontinuous, which permitted a fairly obvious subdivision of patients into two categories. Eight of the 16 patients had RE scores less than 50% correct, and the other eight had RE scores that exceeded 75% correct. The anatomical significance of that subdivision of patients was inferred from CT scans. For patients with RE scores greater than 75%, the primary auditory cortex and the geniculo-temporal pathway appeared to be spared. Patients with RE scores less than 50% showed evidence of significant damage to the primary auditory cortex, and when the scores were less than 30% the damage appeared to extend posteriorly and superiorly into the parietal lobe.

TABLE 2. Stimuli included in each of the five dichotic tests.

Test	*Stimuli*
Digits	one, two, three, four, five, six
High-contrast words	pie, tree, cloud, book, door, glove
Vowel words	key, cow, car, bow, boy, bear
Consonant words	pan, fan, man, boat, coat, goat
CV syllables	/pa/, /ta/, /ka/, /ba/, /da/, /ga/

Tentatively, it is concluded that the dichotic digits test is adequately sensitive to the presence of a central lesion and if the preliminary findings can be confirmed, the digits test may be additionally helpful with respect to identification of lesion location.

What, then, about the question of a coexisting peripheral hearing loss? This question has been addressed in an experiment in which four of the tests described previously were administered to 27 patients with sensorineural hearing loss and for whom there was no reason to believe that a central deficit existed. The high-contrast word test was omitted. The results are shown in Table 3. In this case, because the tests are intended to tap central deficit, a test insensitive to the presence of peripheral hearing loss was desired. By empirically

TABLE 3. Mean, standard deviation, and range of LE and RE scores and of ear advantage (EA) for 27 patients with sensorineural hearing loss.

Test		*LE*	*RE*	*EA*
Digits	Mean	97.2	98.1	0.9
	Sigma	3.7	2.7	2.4
	Range	84/100	92/100	−3/+8
Vowel words	Mean	89.0	94.6	5.6
	Sigma	15.0	7.9	12.7
	Range	50/100	73/100	−18/+48
Consonant words	Mean	81.0	86.7	5.8
	Sigma	11.3	9.5	13.7
	Range	53/99	69/100	−22/+45
CV syllables	Mean	57.0	68.8	11.8
	Sigma	13.4	11.6	16.4
	Range	38/92	48/92	−26/+48

defining insensitivity as a low mean and low standard deviation for the ear advantage, then, digits appear to be most promising.

In summary, from the two experiments described, the dichotic digits test appears to be an acceptable clinical speech test of central auditory function.

Choice of Test Intensity

When the patient for whom a central test is to be administered also has a peripheral hearing loss, choice of test intensity becomes an important issue. There appear to be two approaches that are common among clinicians. One, if threshold sensitivity is different for the two ears, equal SL rather than equal SPL may be adopted. However, no empirical evidence indicates that SL is more appropriate than SPL. Two, if the syllables are presented dichotically with equal SPL, and if the tests intensity corresponds to an intensity for which monotic recognition scores are equal for the two ears, the peripheral loss should not affect the dichotic ear advantage.

Finally, one must consider that equality of monotic recognition scores is necessary, but it is not sufficient to ensure that the dichotic ear advantage is relatively independent of the hearing loss. To support this contention, the following description is offered of results obtained from two hearing-loss patients and from one listener for whom a conductive hearing loss was simulated.

Findings for one patient (Speaks, 1980) are shown in Figure 2. The audiogram is shown at the left, and results for monotic and dichotic recognition of CV nonsense syllables are shown at the right. Monotic recognition scores are shown at the bottom of the right-hand figure, and the point to be made is that monotic scores were essentially equal for the two ears at four intensities: 60, 70, 80, and 90 dB. Dichotic ear advantages are shown at the top with the number reflecting the size of advantage and algebraic sign reflecting the direction of advantage (− for an LEA and + for an REA). At those same four intensities where monotic scores were equal, the dichotic ear advantage ranged from an LEA of 20% at 80 dB to REAs greater than 40% at 60 and 70 dB. Results for a second patient (Speaks, Blecha, & Schilling, 1980) are shown in Figure 3, and the findings are quite similar. Monotic scores were equal at 80, 90 and 100 dB SPL. The dichotic ear advantage, however, was an REA of 16% at 100 dB, no ear advantage at 90 dB, and an LEA of 27% at 80 dB. For these listeners, the direction and size of the ear advantage appeared to reflect the proximity of test intensity to the "knee" of the monotic recognition functions. The problem, of course, is that with these patients there was no knowledge of the dichotic ear advantage that would have been observed had there been no hearing loss.

In order to deal with this problem, a final experiment is described in which a conductive hearing loss was simulated in a single listener by insertion of an EAR plug (Speaks, Bauer, & Carlstrom, 1983). The listener was tested first with no plug

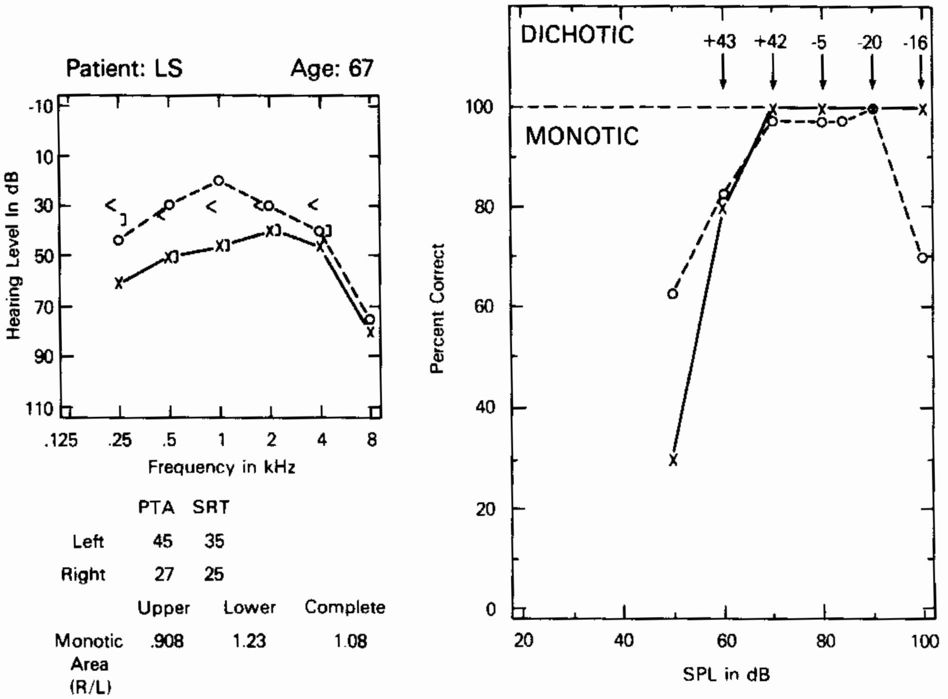

FIGURE 2. Dichotic listening results with CV nonsense syllables for a hearing-impaired patient. The left panel shows results of audiometric testing with pure tones. The right panel shows performance for CV syllables, the lower section shows monotic performance for each ear and the upper section shows percent ear advantage for each of several test intensities.

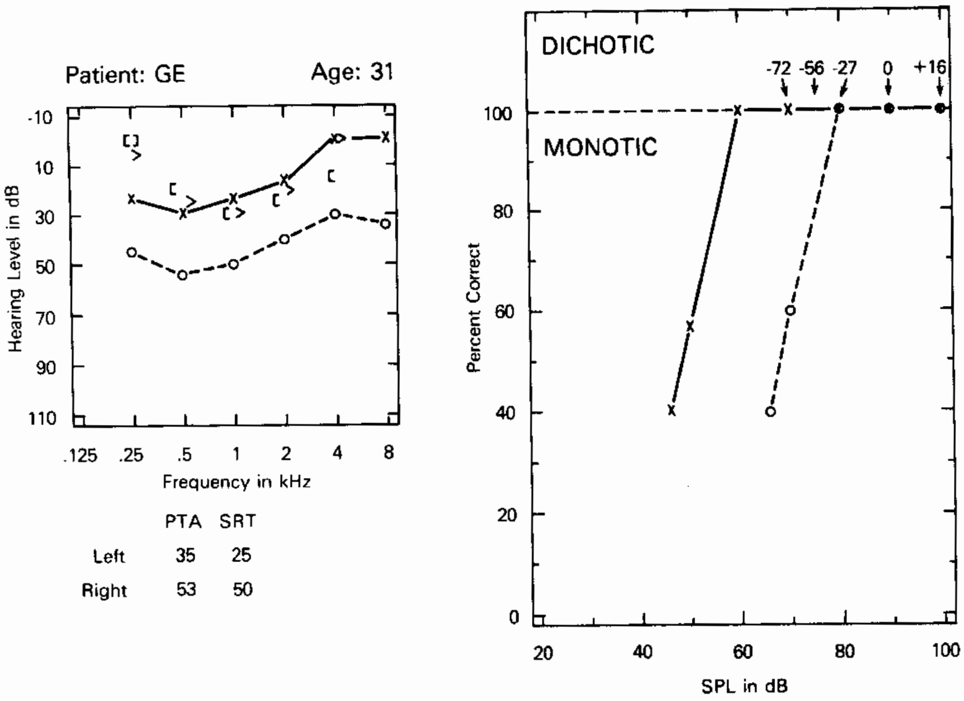

FIGURE 3. Left, results of audiometric testing on a single patient; right, monotic performance-intensity functions for both ears and results of dichotic testing at each of five intensities.

inserted. Monotic recognition functions were defined for both ears, and dichotic ear advantages then were assessed at eight intensities where the monotic scores are essentially equal for the two ears. The results are shown in Figure 4. Only one of the advantages was significant (−7% at 70 dB SPL), and it seems reasonable to say that the ear advantages were essentially independent of test intensity in this case. Figure 5 shows the results that were obtained when a plug was inserted in the left ear (top) and the right ear (bottom). With the plug inserted, a unilateral hearing loss of approximately 35 dB was simulated. In these cases, both the magnitude and direction of percent ear advantage varied with test intensity, even though monotic recognition scores exceeded 95% for both ears. At the top where the left ear was plugged, nonsignificant ear advantages were obtained between 90 and 100 dB, but as test intensity approached the knee of the monotic function for the plugged left ear, the advantage became an REA as large as 42%. At the bottom, nonsignificant ear advantages occurred only in the middle, around 90 dB SPL. As test intensity was lowered toward the knee of the plugged right ear, larger and larger LEAs were observed. In addition, as test intensity was increased above 90 and approached the upper knee of the left ear, the ear advantage reversed to become a large REA.

A few implications of these findings may be listed as they apply to clinical testing (Speaks, Bauer, & Carlstrom, 1983). First, it is imperative that monotic recognition be assessed before administration of a dichotic speech test. Second, it apparently is necessary to define monotic recognition at several intensities in order to gain some indication of the knees of the

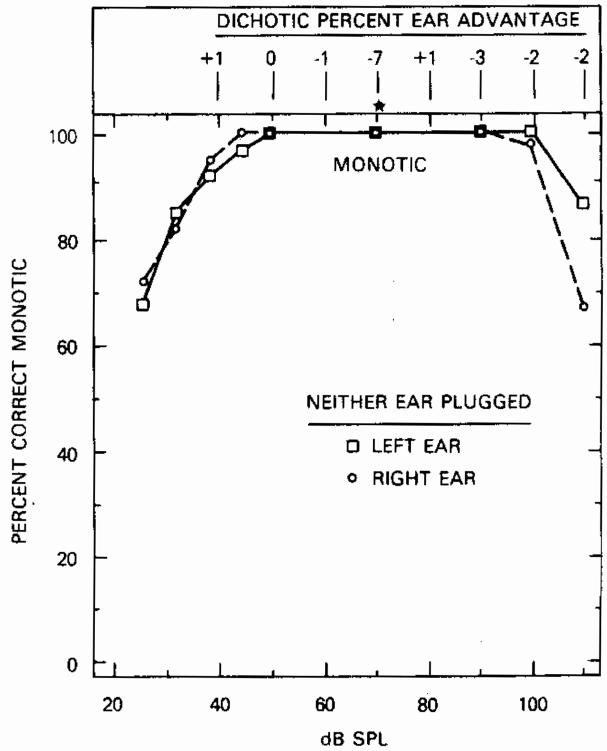

FIGURE 4. Monotic recognition functions for one listener with no plug inserted.

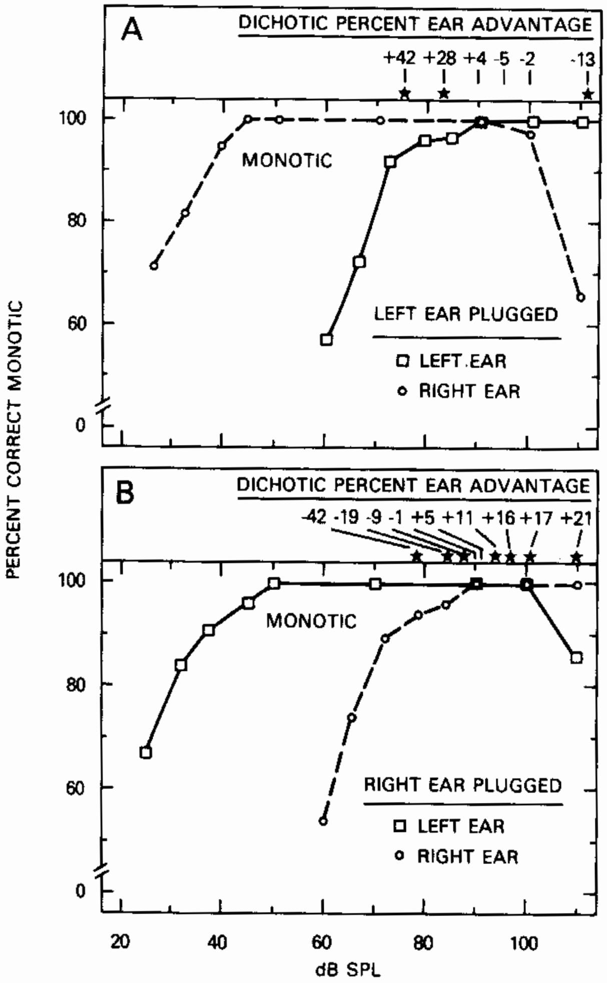

FIGURE 5. Monotic recognition functions for same listener in Figure 4 with left ear plugged (upper panel) and right ear plugged (lower panel).

recognition functions. Third, when CV nonsense syllables are used, the intensity for dichotic testing must be at least 10 dB from both the lower and upper monotic knees for the same nonsense syllables.

Although digits may be a more acceptable central test than CV nonsense syllables, it is not yet known if they are influenced by test intensity in the same way and to the same degree that have been observed for the syllables.

ACKNOWLEDGMENTS

Research was supported by PHS Grant NS-12125 and Contract N01-NS-2378.

REFERENCES

BLUMSTEIN, S., GOODGLASS, H., & TARTTER, V. (1975). The reliability of ear advantage in dichotic listening. *Brain and Language, 2,* 226–236.

CATLIN, J., VAN DERVEER, M. J., & TEICHER, E. D. (1976). Monaural right-ear advantage in a target-identification task. *Brain and Language, 3,* 470–481.

DAMASIO, H., & DAMASIO, A. (1979). "Paradoxic" ear extinction in dichotic listening: Possible anatomic significance. *Neurology, 29,* 644–653.

HAYDEN, M. E., KIRSTEIN, E., & SINGH, S. (1979). Role of distinctive features in dichotic perception of 21 English consonants. *Journal of the Acoustic Society of America, 65,* 1039–1046.

JOHNSON, J. P., SOMMERS, R. K., & WEIDNER, W. E. (1977). Dichotic ear preference in aphasia. *Journal of Speech and Hearing Research, 20,* 116–129.

KATSUKI, J., SPEAKS, C., PENNER, F. & BILGER, R. C. (1984). Application of theory of signal detection to dichotic listening. *Journal of Speech and Hearing Research, 27,* 444–448.

LAUTER, J. L. (1982). Dichotic identification of complex sounds: Absolute and relative ear advantages. *Journal of the Acoustic Society of America, 71,* 701–707.

LINEBAUGH, C. W. (1978). Dichotic ear preference in aphasia: Another view. *Journal of Speech and Hearing Research, 21,* 598–600.

MOORE, W. H., & WEIDNER, W. E. (1975). Dichotic word perception of aphasic and normal subjects. *Perceptual and Motor Skills, 40,* 379–386.

NICCUM, N., RUBENS, A. B., & SPEAKS, C. (1981). Effects of stimulus material on the dichotic listening performance of aphasic patients, *Journal of Speech and Hearing Research, 24,* 526–534.

PETTIT, J. M., & NOLL, J. D. (1979). Cerebral dominance in aphasia recovery. *Brain and Language, 7,* 191–200.

PIZZAMIGLIO, L., DEPASCALIS, C., & VIGNATI, A. (1974). Stability of dichotic listening test. *Cortex, 10,* 203–205.

REPP, B. (1977). Measuring laterality effects in dichotic listening. *Journal of the Acoustical Society of America, 42,* 720–737.

RYAN, W., & MCNEIL, M. (1974). Listener reliability for a dichotic task. *Journal of the Acoustical Society of America, 56,* 1922–1923.

SCHULHOFF, C., & GOODGLASS, H. (1969). Dichotic listening, side of brain injury and cerebral dominance. *Neuropsychologia, 7,* 149–160.

SPARKS, R., & GESCHWIND, N. (1968). Dichotic listening in man after section of neocortical commissures. *Cortex, 4,* 3–16.

SPARKS, R., GOODGLASS, H., & NICKEL, B. (1970). Ipsilateral versus contralateral extinction in dichotic listening resulting from hemispheric lesions. *Cortex, 6,* 249–260.

SPEAKS, C. (1980). Evaluation of disorders of the central auditory system. In M. Paparella & D. Shumrick (Eds.), *Otolaryngology (1846–1860).* Philadelphia: Saunders.

SPEAKS, C., BAUER, K., & CARLSTROM, J. (1983). Peripheral hearing loss: Implications for clinical dichotic listening tests. *Journal of Speech and Hearing Disorders, 48,* 135–139.

SPEAKS, C., BLECHA, M., & SCHILLING, M. (1980). Contributions of monotic intelligibility to dichotic performance. *Ear and Hearing, 1,* 259–266.

SPEAKS, C., & CICCARELLI, T. (1975). *Selective attention in dichotic listening.* Unpublished manuscript.

SPEAKS, C., & NICCUM, N. (1977). Variability of the ear advantage in dichotic listening. *Journal of the American Audiological Society, 3,* 52–57.

SPEAKS C., NICCUM, N., & CARNEY, E. (1982). Statistical properties of responses to dichotic listening with CV nonsense syllables. *Journal of the Acoustical Society of America, 72,* 1185–1194.

WEXLER, E., HALWES, T., & HENINGER, G. (1981). Use of statistical significance criterion in drawing inferences about hemispheric dominance for language function from dichotic listening data. *Brain and Language, 13,* 13–18.

Chapter 14

RECOMMENDATIONS FOR SPEECH RECOGNITION RESEARCH

EARLEEN ELKINS

National Institute of Neurological and Communicative Disorders and Stroke
Bethesda, MD

This report has been an attempt to view speech recognition for the hearing-impaired from a broad perspective. The term *speech recognition* rather than *speech discrimination* has been used throughout this report to reflect current usage and to avoid confusion with a psychophysical task in which the subject is required to compare or differentiate among two or more stimuli. Speech recognition testing has been used in the generic sense to include measures to predict how well a person will function in everyday life, to diagnose hearing impairment, to select among several hearing aids, and to determine how well the message is perceived at the cortical level. Each chapter was selected to either evaluate current clinical usage of speech recognition measures, identify means of improving such measures, or suggest avenues for future investigations. Unanimity was seldom achieved on the various issues discussed. This chapter separates arbitrarily those recommendations into (a) stimulus materials for speech recognition testing; (b) development of speech recognition testing; and (c) test procedures.

STIMULUS MATERIALS

Conventional speech recognition material has consisted primarily of monosyllabic words as witnessed by the use of the PB-50 word lists (Egan, 1948), W-22 lists (Hirsh et al., 1952), and CNC or NU-6 lists (Lehiste & Peterson, 1959; Peterson & Lehiste, 1962; Tillman & Carhart, 1966; Tillman, Carhart, & Wilbur, 1963). Psychometricians and others point out that these word lists do not sample the domain of all speech but only monosyllables and that polysyllables, sentences, and continuous discourse introduce a greater variety of speech syntax and morphology for speech recognition tests. Conference participants felt that patients maintain their relative position on tests using different types of speech stimuli except those that are highly selective of specific speech sounds, such as the California Consonant Test (CCT—Owens & Schubert, 1977) and the Pascoe High Frequency Word Test (Pascoe, 1975).

Some discussion revolved about new or different test materials but did not rule out continued use of current tests. It was strongly reinforced that the test lists are the recorded version only and not a printed list of words or sentences. Any recorded version should be a sample of normal, everyday speech rather than idealized speech, as some recordings of speech recognition tests have been judged. Among newer materials identified as having the potential for obtaining more information about the patient's ability to recognize speech were the Speech Perception in Noise (SPIN) Test (Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984) and the City University of New York Nonsense Syllable Test (NST—Resnick, Dubno, Hoffnung, & Levitt, 1975). The SPIN Test research has shown that the CNC monosyllable stimuli permit the largest number of vowel distinctions, but the CCNC words were more likely to discriminate between listeners with normal hearing and hearing-impaired listeners. An important advantage of the NST is its repeatability and precision similar to that obtained with nonspeech stimuli such as pure tones. Nonsense syllable materials in the form of CV or VC tend to have shorter learning effects. Though some work has been done with children using the NST, different stimulus lists need to be developed for children under 6 years of age to reflect their pattern of errors in phoneme recognition. Suggestions for such an NST included consideration of the use of a carrier phrase, procedures for level equalization, techniques for stimulus presentation, and response elicitation.

Nevertheless, more experimental studies should be encouraged which would vary homogeneity of syntactic structure, communicative import, and other factors from one condition to the next. Materials which incorporate some of the insights of recent speech and language research may prove to be better predictors of real communicative performance. Alternatives to real speech stimuli were discussed, but the majority of the participants felt that the biological relevance of speech makes it the most practical measure of how well a person actually functions in everyday living. However, synthetic speech was suggested as possible stimuli to reduce the source of variability associated with taped representations of speech materials and to permit control of acoustic cues that are considered important.

There still appears to be a need for a number of different tests of speech recognition, depending on how the resulting information is to be used for describing a patient. Tests that are broad tend to predict many attributes a small amount, whereas tests that are very specific, such as the CCT, tend to predict one attribute very well—that is, the recognition of consonants loaded with high-frequency acoustic information. Because different types of hearing-impairment affect recognition of speech cues in different ways, word lists tend to differ in relative difficulty among various types of impairment. Mention also should be made of the need to select stimuli for speech recognition tests that can assess how well the brain receives the message from a damaged peripheral organ.

TEST DEVELOPMENT

Since early tests of speech recognition evolved fro̧m materials and procedures designed to evaluate communication systems (Egan, 1948), it is not surprising that these stimuli have been able to be rather precisely defined by their physical properties. With present-day technology, the acoustical characteristics can be even better described, but many of the fundamental concepts of psychometric measurement have been ignored by most developers of speech recognition tests. For example, few investigators explicitly state the purpose for which the test is recommended. With the exceptions of the Pascoe High Frequency Test, the California Consonant Test, and the SPIN Test (Bilger et al., 1984), the clinician is led to believe that tests of speech recognition are equally valid predictors for indicating degree of hearing handicap, diagnosing site of lesion, or differentiating among appropriate hearing aids. Psychometric theory states that when a group of test items are used for alternate purposes, they cease to be the same test and require different instructions, normative data, and reliability and validity studies. Validity studies can be as diverse as the possible reasons for using the test and the possible inferences that will be drawn about the resulting test scores.

Another shortcoming identified by the workshop participants is the lack of appropriate normative data obtained with speech recognition tests. Two types of norms may be required: those obtained from normal-hearing listeners and those obtained from a specific group such as the hearing impaired or children. Failure to obtain normative data for a test on a representative sample of the people to whom it will be applied has been a common fault among speech recognition test developers. For example, speech-level distributions as published for a group of normal-hearing listeners are unlikely to be appropriate for hearing-impaired listeners who display a reduced dynamic range. Another argument for not norming a clinical test on a population of normal-hearing listeners has to do with reliability. Since reliability for test standards is measured by correlation, it will be high when individual differences are real and when imprecision or unsystematic sources of variation are small.

Ways to improve the reliability of a test were also discussed. The test developer may simply increase the number of items or the amount of information per item. Should clinicians reject the idea of devoting more time to speech recognition testing, they are reminded that future tests, as well as modifications of current tests, will probably be designed to employ computer-assisted techniques and will actually require less testing time.

The need for age-related norms received considerable attention among the discussants. It was generally agreed that they are necessary for children in order to account for their cognitive development and vocabulary. Additionally, each suggestion proposed here for tests of speech recognition by adults is applicable to the development of tests designed to tap a child's ability to process speech signals. It was suggested that only modifications of test methods—that is, allowing more response time—are required for tests of the aged rather than a specific set of norms.

More sensitive diagnostic speech recognition tests are needed which match the particular acoustic characteristics of speech sounds and specific properties of the damaged peripheral mechanism plus the information that it transmits to the central auditory mechanism. Greater cognizance of the acoustics of the speech sounds used in tests of speech recognition should permit the use of Articulation Theory models for predicting speech recognition ability of the listener.

Generally, the workshop participants felt there will never be a perfect test because so many variables, such as filtering, noise, visual cues, and so on, must be considered. Nevertheless, it is incumbent upon all test developers to attempt a very close approximation to a good test by finding the variables that make the largest contribution to the factor of interest and to incorporate them according to accepted psychometric practice.

TEST PROCEDURES

The workshop participants agreed that greater attention needs to be given to statistical assumptions underlying speech recognition test procedures and interpretation of the resulting data. Discussion centered on instruction of the listener, the type of response requested, the effects of computer-assisted techniques, and interpretation of scores.

Accepted psychometric theory requires that the response instructions be adherred to since any deviation constitutes a different test. For example, if "no response" is acceptable, the test cannot be considered the same as when the patient is instructed to guess, if necessary, to provide a response to each test item. The manual for the Revised SPIN Test (Bilger, 1984) states explicitly that "the audiologist should make a special effort to ensure that the client understands that she/he is to repeat the last word of each sentence and that she/he must give a response to each test item" (p. 35). To encourage a response, the manual instructs the audiologist to stop the tape recorder and wait for it. Several of the participants felt that "no response" underestimates the sensory capacity of the listener. It was also suggested that if "no response" occurs more frequently than would be expected, the test administration should be stopped and the patient be reinstructed.

The standard repeat-back response mode was criticized for not reflecting real-life communication situations which seldom, if ever, require it. Perhaps more effective methodologies could be developed for use with more realistic stimulus materials such as sentences and continuous or interactive discourse. Use of a tracking procedure might be appropriate in this context, but like any other instruction strategy, it needs to be validated for actual speech materials.

Another area identified for further research was the selection of a response format within which a child can respond consistently and effectively. The role of cognitive processes in determining performance of young children on word-recognition/picture-recognition tests has not been examined.

Computer-assisted techniques for speech recognition testing are on the horizon and will probably be accepted by progressive clinicians as a savings in both time and energy. Most proponents of computer-assisted audiometry of any type feel that the cost of the equipment will be easily offset by the decrease in personnel and required testing time. Computers will facilitate the use of more standard materials and automatic scoring and will no longer restrict the presentation of test items in serial order as on an audio tape. Thus, logical decisions will be able to be made on the basis of a correct or incorrect response by the patient, and adaptive testing strategies will be permitted. The net result will most likely be a decrease in testing time since the test protocol would use only those items that discriminate the performance of the listener.

With the availability of computers, speech can now be processed in real time so that computer simulation of hearing aids is possible. It will be the responsibility of programmers to tap this capability so that speech recognition tests for hearing aid selection can be utilized to their fullest extent. The ease of multiple calculations can provide better predictors for individualized hearing-aid fittings by using comfortable listening level, dynamic range, sensitivity thresholds, and selected spectral considerations rather than the phonetic and linguistic constraints of the stimulus materials.

Finally, a suggestion was made regarding an improvement in test scoring. When "no response" is acceptable, two scores may be more descriptive of the listener's performance. One would be the conventional percentage based on the total number of test items and the second score would be based on the number of correct and incorrect responses only. This latter score would provide the basis for determining the need and direction for aural rehabilitation.

REFERENCES

BILGER, R. C. (1984) *Manual for the clinical use of the Revised SPIN Test.* Champaign-Urbana: University of Illinois.

BILGER, R. C., NUETZEL, J. M., RABINOWITZ, W. M., & RZECZKOWSKI, C. (1984). Standardization of a test of Speech Perception in Noise. *Journal of Speech and Hearing Research, 27,* 32–48.

EGAN, J. P. (1948). Articulation testing methods. *Laryngoscope, 58,* 955–991.

HIRSH, I., DAVIS, H., SILVERMAN, S. R., REYNOLDS, E., ELDERT, E., & BENSEN, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17,* 321–337.

LEHISTE, I., & PETERSON, G. E. (1959). Linguistic considerations in the study of speech intelligibility. *Journal of the Acoustical Society of America, 31,* 280–286.

OWENS, E., & SCHUBERT, E. D. (1977). Development of the California Consonant Test. *Journal of Speech and Hearing Research, 20,* 463–474.

PASCOE, D. P. (1975). Frequency responses of hearing aids and their effects on the speech perception of hearing-impaired subjects. *Annals of Otology, Rhinology, and Laryngology, 84*(Suppl. 23).

PETERSON, G. E., & LEHISTE, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders, 27,* 62–70.

RESNICK, S. B., DUBNO, J. R., HOFFNUNG, S., & LEVITT, H. (1975). Phoneme errors on a nonsense syllable test. *Journal of the Acoustical Society of America,* 58(Suppl. 1), S114.

TILLMAN, T. W., & CARHART, R. (1966). *An expanded test for speech discrimination utilizing CNC monosyllabic words (Northwestern University Auditory Test No. 6)* (Tech. Rep. SAM-TR-66-55). Brooks AFB, TX: USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC).

TILLMAN, T. W., CARHART, R., & WILBER, L. (1963). *A test for speech discrimination composed of CNC monosyllabic words (Northwestern University Auditory Test No. 4)* (Tech. Rep. SAM-TDR-62-135). Brooks AFB, TX: USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC).

BIBLIOGRAPHY

Abbs, M. S., & Minifie, F. D. (1969). Effect of acoustic cues on fricatives on perceptual confusions in preschool children. *Journal of the Acoustical Society of America, 46,* 1535–1542.

American National Standards Institute. (1960). *American National Standard measurement for monosyllabic word intelligibility* (ANSI S3.2-1960). New York: ANSI.

American Psychological Association. (1974). *Standards for educational and psychological tests.* Washington, DC: APA.

American Speech and Hearing Association, Committee on Rehabilitative Audiology. (1974). The audiologist: Responsibilities in the habilitation of the auditorily handicapped. *Asha, 16,* 14–18.

Beattie, R. C., & Edgerton, B. J. (1976). Reliability of monosyllabic discrimination tests in white noise for differentiating among hearing aids. *Journal of Speech and Hearing Disorders, 41,* 464–476.

Bernstein, L. E. (1979). Developmental differences in labeling VOT continua with varied fundamental frequency. *Journal of the Acoustical Society of America, 65*(Suppl. 1), S1.

Bernstein, L. E. (1982). Ontogenetic changes in children's speech-sound perception. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice.* New York: Academic Press.

Bess, F. H., & Gibler, A. M. (1981). Syllable recognition skills of unilaterally hearing-impaired children. *Asha, 23,* 724.

Bess, F. H., Josey, A. F., & Humes, L. E. (1979). Performance intensity functions in cochlear and eighth nerve disorders. *American Journal of Otolaryngology, 1,* 27–31.

Beving, B., & Eblen, R. (1973). Same and different concepts and children's performance on speech discrimination. *Journal of Speech and Hearing Research, 16,* 513–517.

Bilger, R. C. (1984). *Manual for the clinical use of the Revised SPIN Test.* Champaign-Urbana: University of Illinois.

Bilger, R. C., & Hirsh, I. J. (1956). Masking of tones by bands of noise. *Journal of the Acoustical Society of America, 28,* 623–630.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of Speech Perception in Noise. *Journal of Speech and Hearing Research, 27,* 32–48.

Bilger, R. C., Nuetzel, J. M., Trahiotis, C., & Rabinowitz, W. M. (1980). An objective psychophysical approach to measuring hearing for speech. *Asha, 22,* 726.

Bilger, R. C., & Wang, M. D. (1976). Consonant confusion in patients with sensorineural hearing loss. *Journal of Speech and Hearing Research, 19,* 718–748.

Blumstein, S., Goodglass, H., & Tartter, V. (1975). The reliability of ear advantage in dichotic listening. *Brain and Language, 2,* 226–236.

Braida, L. D., Durlach, N. I., Lippmann, R. P., Hicks, B. L., Rabinowitz, W. M., & Reed, C. M. (1979). *Hearing aids: A review of past research on linear amplification, amplitude compression, and frequency lowering* (ASHA Monographs No. 19). Rockville, MD: American Speech-Language-Hearing Association.

Brandes, P. J., & Ehinger, D. M. (1981). The effects of early middle ear pathology on auditory perception and academic achievement. *Journal of Speech and Hearing Disorders, 46,* 301–307.

Byrne, D. J. (1976). The speech spectrum—Some aspects of its significance for hearing aid selection and evaluation. *British Journal of Audiology, 11,* 40–46.

Castle, W. E. (1963). Effects of selective narrow-band filtering on the perception by normal listeners of Harvard PB-50 Word Lists. *Journal of Speech and Hearing Association of Virginia, 5,* 12–21.

Catlin, J., Van Derveer, M. J., & Teicher, E. D. (1976). Monaural right-ear advantage in a target-identification task. *Brain and Language, 3,* 470–481.

Chen, F. R., Zue, V. W., Picheny, M. A., Durlach, N. I., & Braida, L. D. (1980). Speaking clearly: Acoustic characteristics and intelligibility of stop consonants. *Journal of the Acoustical Society of America, 67*(Suppl. 1), S38.

Chial, M. R., & Hayes, C. S. (1974). Hearing aid evaluation methods: Some underlying assumptions. *Journal of Speech and Hearing Disorders, 39,* 270–279.

Cole, R. A. (Ed.). (1980). *Perception and production of fluent speech.* Hillsdale, NJ: Erlbaum.

Cramer, K. D., & Erber, N. P. (1974). A spondee recognition test for young hearing-impaired children. *Journal of Speech and Hearing Disorders, 39,* 304–311.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronback, L. J., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Damasio, H., & Damasio, A. (1979). "Paradoxic" ear extinction in dichotic listening: Possible anatomic significance. *Neurology, 29,* 644–653.

Danhauer, J. L., & Singh, S. (1975). *Multidimensional speech perception by the hearing impaired.* Baltimore: University Park Press.

Davis, H., & Silverman, S. R. (Eds.). (1970). *Hearing and deafness* (3rd ed.). New York: Holt, Rinehart & Winston.

DeFilippo, C. L., & Scott, B. L. (1978). A method for training and evaluating the reception of ongoing speech. *Journal of the Acoustical Society of America, 63,* 1186–1192.

DeGennaro, S. V. (1978). *The effect of syllabic compression on speech intelligibility for normal listeners with simulated sensorineural hearing loss.* Unpublished master's thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA.

DeGennaro, S. V., Braida, L. D., & Durlach, N. I. (1981). Statistical analysis of third-octave speech amplitude distributions. *Journal of the Acoustical Society of America, 69*(Suppl. 1), S16.

DeRenzi, E., & Vignollo, L. A. (1962). The Token Test: A sensitive test to detect receptive disturbance in aphasics. *Brain, 85,* 665–678.

Dillon, H. (1982). The effect of response methods on the difficulty of speech discrimination tests, A response to Wilson and Antablin, JSHD 1980. *Journal of Speech and Hearing Disorders, 47,* 110–111.

Dirks, D., Kamm, C., Bower, D., & Betsworth, A. (1977). Use of performance-intensity functions for diagnosis. *Journal of Speech and Hearing Disorders, 48,* 408–415.

Dixon, W. J., & Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the Acoustical Society of America, 43,* 109–126.

Dubno, J. R., & Dirks, D. D. (1982). Evaluation of hearing-impaired listeners using a Nonsense Syllable Test. I. Test reliability. *Journal of Speech and Hearing Research, 25,* 135–141.

Dubno, J. R., Dirks, D. D., & Langhofer, L. R. (1982). Evaluation of hearing-impaired listeners using a Nonsense Syllable Test. II. Syllable recognition and consonant confusion patterns. *Journal of Speech and Hearing Research, 25,* 141–148.

Dubno, J. R., & Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *Journal of the Acoustical Society of America, 69,* 249–261.

Dugal, R. L., Braida, L. D., & Durlach, N. I. (1980). Implications of previous research for the selection of frequency-gain characteristics. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance and measurement* (pp. 379–403). Baltimore: University Park Press.

Dunn, H. K., & White, S. D. (1940). Statistical measurements on conversational speech. *Journal of the Acoustical Society of America, 11*, 278–288.

Dunn, L. M. (1981). *Peabody Picture Vocabulary Test.* Circle Pines, MN: American Guidance Service.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope, 58*, 955–981.

Eimas, P. D. (1974). Linguistic processing of speech by young infants. In R. L. Schiefelbusch & L. L. Lloyd (Eds.), *Language perspectives acquisition, retardation and intervention.* Baltimore: University Park Press.

Elliott, L. L. (1979). Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability. *Journal of the Acoustical Society of America, 66*, 651–653.

Elliott, L. L., & Katz, D. R. (1980a). *Northwestern University Children's Perception of Speech (NU-CHIPS): Technical manual.* St. Louis: Auditec.

Elliott, L. L., & Katz, D. R. (1980b). *Development of a new children's test of speech discrimination.* St. Louis: Auditec.

Elliott, L. L., Connors, S., Kille, E., Levin, S., Ball, K., & Katz, D. (1979). Children's understanding of monosyllabic nouns in quiet and noise. *Journal of the Acoustical Society of America, 66*, 12–21.

Elliott, L. L., Longinotti, C., Meyer, D., Raz, I., & Zucker, K. (1981). Developmental differences in identifying and discriminating CV syllables. *Journal of the Acoustical Society of America, 70*, 669–677.

Erber, N. P. (1980). Use of the auditory numbers test to evaluate speech perception abilities of hearing-impaired children. *Journal of Speech and Hearing Disorders, 44*, 527–532.

Fairbanks, G. (1958). Test of phonemic differentiation: The Rhyme Test. *Journal of the Acoustical Society of America, 30*, 596–600.

Fant, G. (1973). *Speech sounds and features.* Cambridge: MIT Press.

Finitzo-Hieber, T., Gerling, I. J., Matkin, N. D., & Cherow-Skalka, E. (1980). A sound effects recognition test for the pediatric audiological evaluation. *Ear and Hearing, 1*, 271–276.

Fitch, H. (1981). Distinguishing temporal information for speaking rate from temporal information for intervocalic stop voicing. *Haskins Laboratories Status Reports, SR-65*, 1–32.

Fletcher, H. (1929). *Speech and hearing* (1st ed.). New York: Van Nostrand.

Fletcher, H. (1953). *Speech and hearing in communication.* New York: Van Nostrand.

Fletcher, H., & Galt, R. H. (1950). Perception of speech and its relation to telephony. *Journal of the Acoustical Society of America, 22*, 89–151.

Fletcher, H., & Steinberg, J. C. (1929). Articulation testing methods. *Bell System Technical Journal, 8*, 806–854.

Flynn, P. T., & Byrne, M. C. (1970). Relationship between reading and selected auditory abilities of third-grade children. *Journal of Speech and Hearing Research, 13*, 725–730.

Franklin, B. (1969). The effect on consonant discrimination of combining a low-frequency passband in one ear with a high-frequency passband in the other ear. *Journal of Auditory Research, 9*, 365–378.

Franklin, B. (1975). The effects of combining low- and high-frequency passbands on consonant recognition in the hearing impaired. *Journal of Speech and Hearing Research, 18*, 719–727.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19*, 90–119.

Giolas, T. G., & Duffy, J. (1973). Equivalency of CID and revised CID sentence lists. *Journal of Speech and Hearing Research, 16*, 739–743.

Giolas, T. G., & Epstein, A. (1963). Comparative intelligibility of word lists and continuous discourse. *Journal of Speech and Hearing Research, 6*, 349–358.

Giolas, T. G. (1966). Comparative intelligibility scores of sentence lists and continuous discourse. *Journal of Auditory Research, 6*, 31–38.

Goetzinger, C. (1972). Word discrimination testing. In J. Katz (Ed.), *Handbook of clinical audiology.* Baltimore: Williams & Wilkins.

Goldman, R., Fristoe, M., & Woodcock, R. (1971). A new dimension in the assessment of speech sound discrimination. *Journal of Learning Disabilities, 4*, 364–368.

Goldman, R., Fristoe, M., & Woodcock, R. (1974). *The Goldman-Fristoe-Woodcock Auditory Skills Test Battery.* Circle Pines, MN: American Guidance Service.

Graham, L. W., & House, A. S. (1971). Phonological oppositions in children: A perceptual study. *Journal of the Acoustical Society of America, 49*, 559–566.

Halle, M. (1977a). Tenseness, vowel shift, and the phonology of the back vowels in modern English. *Linguistic Inquiry, 8*, 611–625.

Halle, M. (1977b). [Review of S. Singh, *Distinctive features: Theory and validation*]. *Journal of the Acoustical Society of America, 62*, 801–802.

Haskins, H. A. (1949). *A phonetically balanced test of speech discrimination for children.* Unpublished master's thesis, Johns Hopkins University, Baltimore.

Hayden, M. E., Kirstein, E., & Singh, S. (1979). Role of distinctive features in dichotic perception of 21 English consonants. *Journal of the Acoustical Society of America, 65*, 1039–1046.

Heidbreder, A. F., & Miller, J. D. (1982). Physical models of a phonetically-relevant auditory-perceptual space. In *Periodic Progress Report No. 25.* St. Louis: Central Institute for the Deaf.

Heisenberg, W. (1976). The nature of elementary particles. *Physics Today, 29*(3), 32–38.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E., Eldert, E., & Bensen, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17*, 321–337.

Hirsh, I., Reynolds, E. G., & Joseph, M. (1954). Intelligibility of different speech materials. *Journal of the Acoustical Society of America, 26*, 530–538.

Hogan, J. T., & Rozsypal, A. J. (1980). Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *Journal of the Acoustical Society of America, 67*, 1764–1771.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. D. (1965). Articulation testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37*, 158–166.

Hudgins, C., Hawkins, J., Karlin, J., & Stevens, S. (1947). The development of recorded auditory tests for measuring hearing loss for speech. *Laryngoscope, 57*, 57–89.

Jakobson, R., & Halle, M. (1956). *Fundamentals of language.* The Hague: Mouton.

Jerger, J., & Jerger, S. (1971). Diagnostic significance of PB word functions. *Archives of Otolaryngology, 93*, 573–580.

Jerger, J., Malmquist, C., & Speaks, C. (1966). Comparison of some speech intelligibility tests in the evaluation of hearing aid performance. *Journal of Speech and Hearing Research, 9*, 253–258.

Johnson, J. P., Sommers, R. K., & Weidner, W. E. (1977). Dichotic ear preference in aphasia. *Journal of Speech and Hearing Research, 20*, 116–129.

Jones, K. O., & Studebaker, G. A. (1974). Performance of severely hearing-impaired children on a closed-response, auditory speech discrimination test. *Journal of Speech and Hearing Research, 47*, 531–540.

Joos, M. (1948). *Acoustical phonetics* (p. 136). Baltimore: Linguistic Society of America.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61*, 1337–1351.

Kamm, C. A., Dirks, D. D., & Carterette, E. C. (1982). Some effects of spectral shaping on recognition of speech by hearing-impaired listeners. *Journal of the Acoustical Society of America, 71*, 1211–1224.

Kamm, C. A., Morgan, D. E., & Dirks, D. D. (1983). Accuracy of adaptive procedure estimates of PB-max level. *Journal of Speech and Hearing Disorders, 48*, 202–209.

Katsuki, J., Speaks, C., Penner, F., & Bilger, R. C. (1984). Application of theory of signal detection to dichotic listening. *Journal of Speech and Hearing Research, 27,* 444–448.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3,* 129–140.

Koike, J. M., & Asp, C. W. (1981). Tennessee Test of Rhythm and Intonation Patterns. *Journal of Speech and Hearing Disorders, 46,* 81–87.

Krause, S. E. (1978). *Developmental use of vowel duration as a cue to postvocalic consonant voicing: A perception and production study.* Unpublished doctoral dissertation, Northwestern University, Chicago.

Kryter, K. D. (1962a). Methods for the calculation and use of the Articulation Index. *Journal of the Acoustical Society of America, 34,* 1689–1697.

Kryter, K. D. (1962b). Validation of the Articulation Index. *Journal of the Acoustical Society of America, 34,* 1698–1702.

Kuhl, P. (1979). The perception of speech in early infancy. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice.* New York: Academic Press.

LaBenz, P. J. (1956). *Potentialities of auditory perception for various levels of hearing loss* (Volta Bureau Reprint 683). Washington, DC: A. G. Bell.

Lauter, J. L. (1982). Dichotic identification of complex sounds: Absolute and relative ear advantages. *Journal of the Acoustical Society of America, 71,* 701–707.

Lehiste, I., & Peterson, G. E. (1959). Linguistic considerations in the study of speech intelligibility. *Journal of the Acoustical Society of America, 31,* 280–286.

Levelt, W. J. (1978). A survey of studies in sentence perception: 1970–1976. In W. J. Levelt & G. M. Flores d'Arcais (Eds.), *Studies in the perception of language.* New York: Wiley.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467–477.

Levitt, H. (1978). Adaptive testing in audiology. *Scandinavian Audiology, 6*(Suppl.), 241–291.

Levitt, H. (1982). *Rehabilitation strategies for the hearing-impaired* (Annual Rep., P01 NS 17764–02). New York: City University of New York.

Levitt, H. (1983). The phoneme: One of life's little uncertainties. In L. Raphael (Ed.), *Language and cognition: Essays in honor of Arthur J. Bronstein.* New York: Plenum.

Levitt, H., Collins, M. J., Dubno, J. R., Resnick, S. B., & White, R. E. C. (1978). *Development of a protocol for the prescriptive fitting of a wearable master hearing aid* (CUNY Research Rep. 11). New York: Communication Science Laboratory.

Levitt, H., & Resnick, S. B. (1978). Speech reception by the hearing-impaired: Methods of testing and the development of new tests. In C. V. Ludvigsen & J. Barford (Eds.), Sensorineural hearing-impairment and hearing aids. *Scandinavian Audiology, 6*(Suppl.), 105–130.

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist, 37,* 148–167.

Linebaugh, C. W. (1978). Dichotic ear preference in aphasia: Another view. *Journal of Speech and Hearing Research, 21,* 598–600.

Linquist, E. (1953). *Educational measurement.* Washington, DC: American Council on Education.

Lippmann, R. (1981). MX41/AR earphone cushions versus a new circumaural mounting. *Journal of the Acoustical Society of America,* 69, 589–592.

Locke, J. L. (1980). The inference of speech perception in the phonologically disordered child. Part II: Some clinically novel procedures, their use, some findings. *Journal of Speech and Hearing Disorders, 45,* 445–468.

Marler, P., & Peters, S. (1981). Birdsong and speech: Evidence for special processing. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives in the study of speech* (pp. 75–112). Hillsdale, NJ: Erlbaum.

Marslan-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science, 189,* 226–228.

Marslan-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10,* 29–63.

McClellan, J. H., Parks, T. W., & Rabiner, L. R. (1973). A computer program for designing optimum FIR linear phase digital filters. *IEEE Transactions on Audio and Electroacoustics, AU-21,* 506–526.

Menary, S., Trehub, S. E., & McNutt, J. (1982). Speech discrimination in preschool children: A comparison of two tasks. *Journal of Speech and Hearing Research, 25,* 202–207.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Miller, J. D. (1981). Predicting aided speech perception. *Journal of the Acoustical Society of America,* 69(Suppl. 1), S98.

Miller, J. D. (1982a, April). Phonetic perception as an auditory-perceptual process. Lecture delivered at *Perception,* a short course of the Office of Continuing Education of the University of Kansas College of Health Sciences and Hospital.

Miller, J. D. (1982b). Auditory-perceptual approaches to phonetic perception. *Journal of the Acoustical Society of America, 71*(Suppl. 1), S112.

Miller, J. D. (1982c). A phonetically-relevant auditory-perceptual space. In *Periodic Progress Report No. 25.* St. Louis: Central Institute for the Deaf.

Miller, J. D. (1982d). An auditory-perceptual approach to phonetic perception. In *Periodic Progress Report No. 25.* St. Louis: Central Institute for the Deaf.

Miller, J. D. (1982e, November). *A phonetically-relevant auditory-perceptual space.* Paper presented at the 104th Meeting of the Acoustical Society of America, Orlando, FL.

Miller, J. D., Engebretson, A. M., Garfield, S. A., & Scott, B. L. (1975). New approach to speech-reception testing. *Journal of the Acoustical Society of America, 57*(Suppl. 1), S48.

Miller, J. D., Engebretson, A. M., & Vemula, N. R. (1982). Observations of the acoustic description of vowels as spoken by children, women, and men. *Journal of the Acoustical Society of America, 68*(Suppl. 1), S33.

Miller, J. D., Engebretson, A. M., & Vemula, N. R. (1980a). Transposition of vowel sounds. In *Periodic Progress Report No. 23.* St. Louis: Central Institute for the Deaf.

Miller, J. D., Engebretson, A. M., & Vemula, N. R. (1980b). Vowel normalization: Differences between vowels spoken by children, women, and men. *Journal of the Acoustical Society of America, 68*(Suppl. 1), S33.

Miller, J. D., Niemoeller, A. F., Pascoe, D. P., & Skinner, M. W. (1980). Integration of the electroacoustic description of hearing aids with the audiologic description of clients. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 355–377). Baltimore: University Park Press.

Milner, P. (1973). Advantages of experienced listeners in intelligibility testing. *IEEE Transactions on Audio and Electroacoustics, AU-21,* 161–165.

Milner, P. (1982). *Perception of filtered speech by hearing-impaired listeners and by normal listeners with simulated hearing loss.* Unpublished doctoral dissertation, City University of New York.

Mills, J. H. (1975). Noise and children: A review of the literature. *Journal of the Acoustical Society of America, 58,* 767–779.

Moore, W. H., & Weidner, W. E. (1975). Dichotic word perception of aphasic and normal subjects. *Perceptual and Motor Skills, 40,* 379–386.

Morse, P. A. (1978). Infant speech perception: Origins, processes, and Alpha Centauri. In F. D. Minifie & L. L. Lloyd (Eds.), *Communicative and cognitive abilities—Early behavioral assessment.* Baltimore: University Park Press.

Morton, J. (1979). Word recognition. In J. C. Marshall & J. Morton (Eds.), *Psycholinguistics 2: Structures and processes.* Cambridge: MIT Press.

Nabelek, I. V., Wood, W. S., & Koike, K. J. M. (1980). Speech perception through various signal processings. *Journal of the Acoustical Society of America, 68*(Suppl. 1), S58.

Niccum, N., Rubens, A. B., & Speaks, C. (1981). Effects of stimulus material on the dichotic listening performance of aphasic patients. *Journal of Speech and Hearing Research, 24,* 526–534.

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Owens, E. (1961). Intelligibility of words varying in familiarity. *Journal of Speech and Hearing Research, 4,* 113–129.

Owens, E., Benedict, M., & Schubert, E. D. (1972). Consonant phonemic errors associated with pure-tone configurations and certain kinds of hearing impairments. *Journal of Speech and Hearing Research, 15,* 308–322.

Owens, E., Kessler, D., & Schubert, E. D. (1982). Interim assessment of candidates for cochlear implants. *Archives of Otolaryngology, 108,* 478–483.

Owens, E., Kessler, D., Telleen, C., & Schubert E. D. (1981, September). The minimal auditory capabilities (MAC) battery. *Hearing Aid Journal, 34,* 9–34.

Owens, E., & Schubert, E. D. (1977). Development of the California Consonant Test. *Journal of Speech and Hearing Research, 20,* 463–474.

Palva, A. C. (1965). Filtered speech audiometry, I. Basic studies with Finnish speech toward the creation of a method for the diagnosis of central auditory disorders. *Acta Otolaryngologica,* (Suppl. 210).

Pascoe, D. P. (1975). Frequency responses of hearing aids and their effects on the speech perception of hearing-impaired subjects. *Annals of Otology, Rhinology, and Laryngology, 84*(Suppl. 23).

Peterson, G. E. (1952). The information bearing elements of speech. *Journal of the Acoustical Society of America, 24,* 629–637.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Peterson, G. E., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders, 27,* 62–70.

Petit, J. M., & Noll, J. D. (1979). Cerebral dominance in aphasia recovery. *Brain and Language, 7,* 191–200.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1980). Speaking clearly: Intelligibility and acoustic characteristics of sentences. *Journal of the Acoustical Society of America, 67*(Suppl. 1), S38.

Pisoni, D. B. (1981a). Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America, 70*(Suppl. 1), S98.

Pizzamiglio, L., DePascalis, C., & Vignati, A. (1974). Stability of dichotic listening test. *Cortex, 10,* 203–205.

Pollack, I. (1948). Effects of high-pass and low-pass filtering on the intelligibility of speech in noise. *Journal of the Acoustical Society of America, 20,* 259–266.

Popelka, G. R., & Engebretson, A. M. (1983). A computer-based system for hearing aid assessment. *Hearing Instruments, 34,* 6–9, 44.

Pols, L. C. W. (1977). *Spectral analysis and identification of Dutch vowels in monosyllabic words.* Soesterberg, Netherlands: Institute for Perception TNO.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals.* Englewood Cliffs, NJ: Prentice-Hall.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the voicing characteristic of the word-final consonants in American English. *Journal of the Acoustical Society of America, 51,* 1296–1303.

Repp, B. (1977). Measuring laterality effects in dichotic listening. *Journal of the Acoustical Society of America, 42,* 720–737.

Resnick, S. B., Dubno, J. R., Hoffnung, S., & Levitt, H. (1975). Phoneme errors on a nonsense syllable test. *Journal of the Acoustical Society of America, 58*(Suppl. 1), S114.

Resnick, S. B., Dubno, J. R., Hawie, D. G., Hoffnung, S., Freeman, L., & Slosberg, R. M. (1976). *Phoneme identification on a closed response nonsense syllable test.* Paper presented at the Annual Convention of the American Speech and Hearing Association, Houston.

Revoile, S., Pickett, J. M., Holden, L., & Talkin, D. (1980). Effects of some acoustic cue modifications on the perception of voiced and unvoiced final stop consonants for hearing-impaired listeners. *Journal of the Acoustical Society of America, 67*(Suppl. 1), S78.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics, 22,* 400–407.

Rosenthal, R. D., Lang, J. K., & Levitt, H. (1975). Speech reception with low frequency speech energy. *Journal of the Acoustical Society of America, 57,* 949–955.

Ross, M., & Lerman, J. W. (1970). A picture identification test for hearing-impaired children. *Journal of Speech and Hearing Research, 13,* 44–53.

Ryan, W., & McNeil, M. (1974). Listener reliability for a dichotic task. *Journal of the Acoustical Society of America, 56,* 1922–1923.

Sanderson-Leepa, M. E., & Rintelmann, W. F. (1976). Articulation functions and test-retest performance of normal-hearing children on three speech discrimination tests: WIPI, PBK-50, and NU Auditory Test No. 6. *Journal of Speech and Hearing Disorders, 41,* 503–519.

Schulhoff, C., & Goodglass, H. (1969). Dichotic listening, side of brain injury and cerebral dominance. *Neuropsychologia, 7,* 149–160.

Schwartz, A. H., & Goldman, R. (1974). Variables influencing performance on speech-sound discrimination tests. *Journal of Speech and Hearing Research, 17,* 25–32.

Schwartz, D. M., & Surr, R. K. (1979). Three experiments on the California Consonant Test. *Journal of Speech and Hearing Disorders, 44,* 61–72.

Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David & P. Denes (Eds.), *Human communication: A unified view* (pp. 67–113). New York: McGraw-Hill.

Sher, A. E., &·Owens, E. (1974). Consonant confusions associated with loss above 2000 Hz. *Journal of Speech and Hearing Research, 17,* 669–681.

Shore, I., Bilger, R., & Hirsh, I. (1960). Hearing aid evaluation: Reliability of repeated measurements. *Journal of Speech and Hearing Disorders, 25,* 152–170.

Siegenthaler, B., & Haspiel, G. (1966). *Development of two standardized measures of hearing for speech by children* (Project No. 2372, Contract No. OE-5-10-003). Washington, DC: U.S. Department of Health, Education and Welfare.

Singh, S., Woods, D. R., & Becker, G. M. (1973). Perceptual structure of 22 prevocalic English consonants. *Journal of the Acoustical Society of America, 52,* 1698–1713.

Skinner, M. W. (1980). Speech intelligibility in noise-induced hearing loss: Effects of high-frequency compensation. *Journal of the Acoustical Society of America, 67,* 306–317.

Skinner, M. W., Karstaedt, M. M., & Miller, J. D. (1982). Amplification bandwidth and speech intelligibility for two listeners with sensorineural hearing loss. *Audiology, 21,* 251–268.

Skinner, M. W., & Miller, J. D. (1983). Amplification bandwidth and intelligibility of speech in quiet and noise for listeners with sensorineural hearing loss. *Audiology, 22,* 253–279.

Skinner, M. W., Pascoe, D. P., Miller, J. D., & Popelka, G. R. (1982). Measurements to determine the optimal placement of speech energy within the listener's auditory area: A basis for selecting amplification characteristics. In G. A. Studebaker & F. H. Bess (Eds.), *The Vanderbilt hearing-aid report* (Monographs in Contemporary Audiology, 161–169). Upper Darby, PA: Associated Hearing Instruments.

Sparks, R., & Geschwind, N. (1968). Dichotic listening in man after section of neocortical commissures. *Cortex, 4,* 3–16.

Sparks, R., Goodglass, H., & Nickel, B. (1970). Ipsilateral versus contralateral extinction in dichotic listening resulting from hemispheric lesions. *Cortex, 6,* 249–260.

Speaks, C. (1980). Evaluation of disorders of the central auditory system. In M. Paparella & D. Shumrich (Eds.), *Otolaryngology (1846–1860).* Philadelphia: Saunders.

Speaks, C. (1967). Intelligibility of filtered synthetic sentences. *Journal of Speech and Hearing Research, 10,* 289–298.

Speaks, C., Bauer, K., & Carlstrom, J. (1983). Peripheral hearing loss: Implications for clinical dichotic listening tests. *Journal of Speech and Hearing Disorders, 48,* 135–139.

Speaks, C., Blecha, M., & Schilling, M. (1980). Contributions of monotic intelligibility to dichotic performance. *Ear and Hearing, 1,* 259–266.

Speaks, C., & Ciccarelli, T. (1975). *Selective attention in dichotic listening.* Unpublished manuscript.

Speaks, C., & Jerger, J. (1965). Methods for measurement of speech identification. *Journal of Speech and Hearing Research, 8,* 185–194.

Speaks, C., Jerger, J., & Trammell, J. (1970). Comparison of sentence identification and conventional speech discrimination scores. *Journal of Speech and Hearing Research, 13,* 755–767.

Speaks, C., & Niccum, N. (1977). Variability of the ear advantage in dichotic listening. *Journal of the American Audiological Society, 3,* 52–57.

Speaks, C., Niccum, N., & Carney, E. (1982). Statistical properties of responses to dichotic listening with CV nonsense syllables. *Journal of the Acoustical Society of America, 72,* 1185–1194.

Stark, R. E., & Tallal, P. (1981). Perceptual and motor deficits in language-impaired children. In R. W. Keith (Ed.), *Central auditory and language disorders in children.* Houston: College-Hill Press.

Stevens, K. N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America, 68,* 836–842.

Stevens, K. N., & Blumstein, S. E. (1980). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. R. Miller (Eds.), *Perspectives on the study of speech* (pp. 1–38). Hillsdale, NJ: Erlbaum.

Stevens, K. N., & Hawkins, S. (1982). Acoustic and perceptual correlates of nasal vowels. *Journal of the Acoustical Society of America, 71*(Suppl. 1), S76.

Studebaker, G. A. (1982). Hearing aid selection: An overview. In G. A. Studebaker & F. H. Bess (Eds.), *The Vanderbilt hearing-aid report* (Monographs in Contemporary Audiology). Upper Darby, PA: Associated Hearing Instruments.

Studebaker, G. A., & Pavlovic, C. V. (1983). A nonsense syllable test designed for articulation index testing. *Journal of the Acoustical Society of America, 73*(Suppl. 1), S102.

Swets, J. A. (Ed.). (1964). *Signal detection and recognition by human observers.* New York: Wiley.

Tallal, P., Stark, R., Kallman, C., & Mellits, D. (1981). A reexamination of some nonverbal perceptual abilities of language-impaired and normal children as a function of age and sensory modality. *Journal of Speech and Hearing Research, 24,* 351–357.

Thomas, I. B., & Pfannebecker, G. B. (1974). Effects of spectral weighting of speech in hearing-impaired subjects. *Journal of the Audio Engineering Society, 22,* 690–694.

Tillman, T. W., & Carhart, R. (1966). *An expanded test for speech discrimination utilizing CNC monosyllabic words* (Northwestern University Auditory Test No. 6) (Tech. Rep. SAM-TR-66-55). Brooks AFB, TX: USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC).

Tillman, T. W., Carhart, R., & Wilber, L. (1963). *A test for speech discrimination composed of CNC monosyllabic words* (Northwestern University Auditory Test No. 4) (Tech. Rep. SAM-TDR-62-135). Brooks, AFB, TX: USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC).

Villchur, E. (1970). Audiometer-earphone mounting to improve intersubject and cushion-fit reliability. *Journal of the Acoustical Society of America, 48,* 1387–1396.

Walden, B. E., Schwartz, D. M., Montgomery, A. A., & Prosek, R. A. (1981). A comparison of the effects of hearing impairment and acoustic filtering on consonant recognition. *Journal of Speech and Hearing Research, 24,* 32–43.

Waltzman, S., & Levitt, H. (1978). The SIL as a predictor of face-to-face communication. *Journal of the Acoustical Society of America, 63,* 581–590.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America, 54,* 1248–1266.

Wang, M. D., Reed, C. M., & Bilger, R. C. (1978). A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusion. *Journal of Speech and Hearing Research, 21,* 5–36.

Weiner, P. S. (1969). The cognitive functioning of language deficient children. *Journal of Speech and Hearing Research, 12,* 53–64.

Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society, B25,* 1–48.

Wetherill, G. B., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology, 18,* 1–10.

Wexler, E., Halwes, T., & Heninger, G. (1981). Use of statistical significance criterion in drawing inferences about hemispheric dominance for language function from dichotic listening data. *Brain and Language, 13,* 13–18.

Williams, C. E., & Hecker, M. H. L. (1968). Relation between intelligibility scores for four test methods and three types of speech distortion. *Journal of the Acoustical Society of America, 44,* 1002–1006.

Wilson, R. H., & Antablin, J. K. (1980). A picture identification task as an estimate of the word-recognition performance of nonverbal adults. *Journal of Speech and Hearing Disorders, 45,* 223–237.

Wilson, R. H., & Antablin, J. K. (1982). The Picture Identification Task, A reply to Dillon. *Journal of Speech and Hearing Disorders, 47,* 111–112.

Winer, B. (1971). *Statistical principles in experimental design.* New York: McGraw-Hill.

Zlatin, M. A., & Koenigsknecht, R. A. (1975). Development of the voicing contrast: Perception of stop consonants. *Journal of Speech and Hearing Research, 18,* 541–553.