Canadian Acoustics / Acoustique Canadienne 24(1) 3-11 (1996) ***Research article / Article de recherche***

DEVELOPMENT, EVALUATION AND SCORING OF A NONSENSE WORD TEST SUITABLE FOR USE WITH SPEAKERS OF CANADIAN ENGLISH

Margaret F. Cheesman and Donald G. Jamieson
Hearing Health Care Research Unit
Department of Communicative Disorders
The University of Western Ontario
London, ON N6G 1H1 CANADA

SUMMARY

Hearing researchers and clinicians frequently need to estimate the overall accuracy of consonant identification for a listener, over time or in various listening conditions, and to know how frequently specific types of consonant confusion errors are made in each condition. The present paper summarizes the development of a closed-set nonsense word test that provides both a general measure of listeners' abilities to identify consonant sounds, and an indication of the types of confusion errors that listeners make. The acoustical characteristics of test items and statistics of performance measures are summarized and two different scoring procedures are evaluated. The test, termed the University of Western Ontario Distinctive Features Differences test (UWODFD), is comprised of high-quality digital recordings of 21 items spoken by four native speakers of Canadian English; two male and two female. All items occur in a fixed, word-medial context. All aspects of testing, including presentation of stimuli, recording of subject responses, and the scoring and presentation of results, are under computer control. The test can be administered relatively quickly, it has been found to be appropriately sensitive to changes in listening conditions and has been used successfully with listeners from a variety of linguistic backgrounds.

SOMMAIRE

Les chercheurs et les practiciens orthophonistes ont souvent besoin d'estimer la totale exactitude de l'idenfication des consonnes, selon un laps de temps précis ou dans des conditions d'écoute variées, et de savoir à quelle fréquence des erreurs de confusion des types spécifiques de consonnes, se produisent dans chaque condition. Cet article résume le développement d'un test sur un ensemble délimité de mots insignificants qui implique à la fois une évaluation générale des capacités de l'auditeur à identifier des sons consonantiques, et une indication des types d'erreurs de confusion que font les auditeurs. Les caractéristiques des tests de mots et les statistiques des degrés de performance sont condensées et deux différentes procédures de scores sont évaluées. Le test, intitulé de Test Différetiel des Traits Distinctifs de l'Université de Western Ontario [UWODFD], est effectué à partir d'enregistrements digitaux de haute qualité de 21 mots énoncés par 4 anglophones canadiens; deux masculins et deux féminins. Tous les phonèmes apparaissent dans un contexte déterminé, en position médiane du mot. Tous les aspects du test, y compris la présentation des stimuli, l'enregistrement des réponses du sujet, le score et la présentation des résultats, sont sous contrôle informatique. Le test peut être administré relativement vite, il s'est avéré suffisamment sensible [de façon appropriée] aux changements des conditions d'écoute et a été utilisé avec succès auprès d'auditeurs de diverses provenances.

1. INTRODUCTION

The measurement of the ability of a listener to perceive spoken language is a fundamental need for many audiologists and hearing researchers. As examples, such measurements are important in medico-legal applications requiring a measurement of the speech-related hearing disability experienced by a listener, in rehabilitation research applications requiring quantification of the benefit that a specific hearing aid provides to a given listener, in the development of improved speech compression, synthesis, and coding systems, and in the evaluation of spoken language learning by English as a

second language students. Suitable speech intelligibility tests must be sensitive (i.e., yielding different results for different listening conditions), valid (i.e., yielding results that are related to "real-world" performance), reliable (i.e., yielding results that are highly reproducible), and feasible (i.e., able to be used easily by subjects and clinicians working under typical circumstances).

No single test is likely to meet all needs, so that a battery of tests is normally required for use in research studies. One important component of such a battery is a measure of listeners' abilities to understand speech based purely on the acoustic information provided to them -- i.e., for which performance was not strongly influenced by higher-level cognitive ability. A test should be computer-controlled, with high-quality recordings of speech in the dialect of the local subject pool. For hearing researchers at the University of Western Ontario, subjects are typically native speakers of central Canadian English. Researchers in this group required a test that provided both an overall measure of intelligibility and a diagnostic measure with which to characterize the specific pattern of confusion errors made by listeners. Unable to find an existing test that met these criteria, a new test was developed which drew from the assets of several existing tests. This paper describes the development and evaluation of this test.

1.1. Objectives and Test Specifications

Five characteristics of the test materials were determined to be essential characteristics of the new test: (1) the target sounds should be representative of all consonant sounds; (2) target sounds should be presented in intervocalic position, to approximate the contextual cues to consonant identity which are available in "running" speech; (3) speech tokens should be obtained from at least four talkers, two men and two women, the accents of all speakers being appropriate for the typical UWO listener; (4) high-quality digitized acoustic signals should be used; (5) all speech tokens should be free of idiosyncracies and anomalies in pronunciation and intonation, and free of apparent accent to central Canadian English-speaking listeners.

Four characteristics of the test implementation were determined to be key: (1) the test was to be automated with stimulus selection, stimulus presentation, presentation of response alternatives, response recording, response scoring, and presentation of results to be under computer control; (2) the administration of a complete form of the test was to require not longer than five minutes, under typical testing situations; (3) the test was to be suitable for use with all adult subjects, and by most adult patients typically seen in clinical situations; and (4) a final characteristic of the test was the ability to analyze the test results in a variety of ways, including overall percentage of correct responses, confusion matrices and feature scoring.

2. SELECTIVE LITERATURE REVIEW

Analytic consonant perception tests have received increasing attention for use in audiological habilitation, particularly as potential tools for hearing aid evaluation (e.g., Jamieson, Brennan & Cornelisse, 1995). Test results can be summarized in the same way as the tests traditionally used as part of an audiological assessment, when they are scored in terms of the overall percent correct word identification or as the signal-to-noise level required to obtain some fixed level of performance. However, in addition, analytic speech perception test responses can be examined with respect to the pattern of errors that occur, i.e. they can provide analytical insight into the nature of the perceptual confusion.. Either confusion matrices (Dubno & Levitt, 1981; Miller & Nicely, 1955; Gordon-Salant, 1987) or feature-based scoring (Feeney & Franks, 1982; Danhauer & Singh, 1975; Miller & Nicely, 1955) can be used to quantify the pattern of response errors.

A number of tests have been developed to address objectives similar to those outlined above. Phoneme-based tests introduced over the past 20 years include the CUNY Nonsense Syllable test (Levitt & Resnick, 1978; Resnick, Dubno, Hoffnung & Levitt, 1975; CUNY-NST), the Modified Rhyme test (House, Williams, Hecker & Kryter, 1965; MRT), the Diagnostic Rhyme test (Voiers, 1983; DRT), and the Four Alternative Auditory Features test (Foster & Haggard, 1987; FAAF). Such tests restrict the set of response alternatives available to the listener on any particular trial to a subset of the complete consonant set. The choice of alternatives is based on the *a priori* probability of errors and/or restriction to confusions along a particular (feature) dimension.

As an example, the CUNY-NST tests initial and final consonant positions separately, within three vowel environments, /i/, /a/, and /u/. It contains 62 items, grouped into seven subtests. Each subtest is designed to measure consonant identification with a focus on a particular feature, within a syllable-initial or syllable-final position, and in one of the three vowel contexts. However, because the testing format involves a restricted set of speech stimuli and possible responses, subjects' confusion errors are restricted to those stimuli contained in the specific distractor set. In many instances, it is of interest to determine which errors subjects will make, when the range of these errors is not constrained through the *a priori* selection of the stimulus and response sets.

2.1. Feature-Based Testing

Some speech testing procedures offer the advantage of permitting feature-based scoring procedures. Feature-based scoring procedures measure performance in terms of a set of acoustic, phonetic, or perceptual features, rather than merely in terms of the proportion of complete consonant targets that are identified correctly. A feature approach has appeal for both clinical and research applications because it may be

more sensitive to small differences in listening conditions than is whole item scoring (Feeney & Franks, 1982). Feature-based testing may therefore permit more efficient assessments of speech perception ability than testing based on whole items.

A similar argument has been made by Boothroyd (1968), who proposed scoring word lists on a phoneme-by-phoneme basis to increase the sensitivity of tests based on word lists. Efficiency is particularly important, because testing is costly and testing time is often severely restricted, such as when speech perception ability is assessed as part of hearing aid evaluation research, or in clinical applications. Historically, the routine application of feature-based scoring procedures has been precluded by the relatively complex scoring methods required. However, the widespread availability of computer-assisted testing protocols in audiological facilities has reduced such considerations.

2.2. The Distinctive Feature Differences Test (DFD)

Feeney and Franks (1982) developed a closed-set consonant recognition task that was designed to be scored on the basis of a set of distinctive feature confusions rather than whole phoneme recognition. This Distinctive Feature Difference (DFD) test was formed from 13 target consonants (/b,t,d,f,dz,k,p,s,ʃ,tʃ,θ,ð,v/) presented in an /ʌCɪl/ context (e.g., "abil"). These 13 consonants were chosen because they were the consonants frequently perceived in error by hearing-impaired listeners, when presented in word-initial or word-final positions (Owens & Schubert, 1968). Because target consonants occurred in syllable-medial position in the DFD test, contextual cues to consonant identity were preserved in adjacent portions of the vowel-consonant-vowel syllable (VCV) as would be expected to occur for many consonants in continuous speech.

Feeney and Franks (1982) reported that feature-based scoring of the DFD test increased the reliability of the speech discrimination scores, because the number of scoreable units in the test could be increased without changing the amount of time required to complete the task. However, reliability coefficients were not reported for their test. Moreover, the DFD test was not automated, complicating administration, data collection, and scoring.

3. PRESENT WORK

The present study describes the development of a DFD test that is automated and consists of high quality digital recordings of all test items. Whole-item scoring for this test has been compared with feature-based scoring and both scoring procedures have been used in a variety of applications.

The test set includes a larger set of test items than Feeney and Franks' (1982) DFD test. For this test, designated the University of Western Ontario DFD (UWODFD), the set of consonant targets was increased to include most single English intervocalic consonants. The UWODFD is essentially an "open-set" test, because it includes most of the single consonants that can occur in the given context. The larger set of consonants allows listeners to make a broad range of perceptual errors and increases the range of perceptual confusions that can occur and the variety of alternative scoring schemes that can be used. To further increase generalizability, four different talkers, two men and two women, were used so the test includes a range of voices and speaking styles. All talkers were native speakers of central Canadian English, thereby increasing the appropriateness of the test for use with an anglophone Canadian subject or client population.

4. STIMULUS PREPARATION

4.1. Test items

Initial target test items were nonsense words of the form /ʌCɪl/ in which **C** was one of the 22 consonants /b, tʃ, d, f, g, h, j, k, l, m, n, p, r, s, ʃ, t, θ, ð, v, w, y, z/ spoken by one of four talkers. The talkers were two male and two female young adults. All were native speakers of central Canadian English.

4.2. Recordings

To obtain the initial set of tokens, each talker was instructed to utter each target token within the carrier phrase "Point to the word /ʌCɪl/". Several tokens of each word were digitized using the carrier phrase, while minimizing variation in the peak levels of the phrase across tokens. All recordings were made with the talker seated in a double-walled, IAC, sound-attenuating room, using a Shure unidirectional microphone coupled to a Shure M267 mixer. The output signal from the mixer was low-pass filtered at 8.0 kHz (Kemo VBF 25MD) and sampled to disk (16-bit recording at 20 kHz via an Ariel DSP-16 A/D card), using the Computerized Speech Research Environment (CSRE) software (Avaaz Innovations, 1995; Jamieson, Ramji, Kheirallah & Nearey, 1992). The test tokens were then edited from the carrier phrase.

4.3. Item selection

A series of behavioural tests was prepared that presented the speech tokens together with a list of the full set of response alternatives displayed on the computer screen (see below). Individual listeners then performed a sequence of tasks to identify speech sounds to be included in the final test protocol. This approach identified speech tokens that met the following criteria: (1) tokens were readily identifiable as the target sounds when presented in quiet to normally-hearing listeners; (2) tokens were rated as good exemplars of the target category; and (3) tokens were determined to be free of idiosyncracies such as atypical pitch contours, loudness differences, or pronunciation irregularities. Tokens that failed to meet all three criteria were deleted from the candidate set.

5. INSTRUMENTATION

Prior to statistical measurement of the long-term spectrum of the stimuli and their subsequent use in the perceptual tests reported here, the 84 digitized stimuli were converted to 12-bit samples to enable the tests to be undertaken with the equipment described below. Stimulus presentation was controlled with a DT-2801A D/A converter and low-pass filtered at 8.0 kHz. Signal level was controlled using a TTE PA-2 programmable attenuator and an Amcron D-75 amplifier.

For behavioural testing, the stimuli were presented monaurally to listeners via TDH-49 earphones. Listeners were tested individually while seated in an IAC double-walled sound-attenuating booth. The masking noise was generated by a TTE white noise generator and shaped to the ⅓-octave band L(eq, 5 min) of the 84 stimuli with two Industrial Research Products DG-4017 equalizers applied in series. The full-band long-term L(eq, 5 min) of the speech-shaped noise was 70 dB(A).

6. PILOT TESTING

Pilot testing with 16 normal-hearing young adult listeners was used to select the final test stimuli from the multiple recordings of each test item. During this testing, subjects were given a list of all recorded test items and were asked to identify each medial consonant when presented at 70 dB SPL. The final test items selected were highly intelligible under such optimal listening conditions, being identified with 95% accuracy or better, and were free of apparent idiosyncracies such as unusual intonation contours or syllable durations that might serve as cues to the identity of the consonant after repeated presentations of the test items.

The nonsense word /ʌθɪl/ was originally included in the set of test items. The results of the pilot identification testing indicated that, despite repeated attempts to obtain highly recognizable test tokens, /θ/ tokens were confused very often with /f/ tokens by the normal hearing listeners in quiet. Furthermore, inclusion of both the voiced and voiceless alveo-dental fricatives /θ/ and /ð/, for which English has no orthographic distinction, required some level of phonetic training and sophistication for the listeners and resulted in response errors that may have not accurately reflected perceptual errors. This is one limitation of a set of test materials that includes a wide variety of possible consonantal responses; the test format must provide unambiguous response items that are constrained by common orthographic practise. Elimination of /ʌθɪl/ resulted in a set of 21 response alternatives that could be unambiguously described using standard English spelling.

7. STATISTICAL DESCRIPTION OF STIMULI

Statistical descriptions of the long-term spectrum of the 84 stimuli (4 talkers x 21 consonants) were obtained through the sound delivery system using a Bruel and Kjaer 2231 sound level meter, statistical module BZ-7101, and a 1625 filter set using ⅓-octave settings. All measurements were made in a 6-cm^3 coupler. Statistical analyses of 5 minute samples of the continuous output (no silent gaps) of the 84 stimuli were made in ⅓-octave bands from 125 to 8000 Hz. The band pressure levels which were exceeded in 1%, 10%, 50%, 90%, and 99% of the 125 ms measurement intervals, and the L(eq) (Earshen, 1986), were measured when the overall level of the speech was adjusted to 70 dB(A).

The distribution of the ⅓-octave long-term speech levels is shown in Figure 1. The spectrum is dominated by the repeated high-intensity portions of the test stimuli, that is, the initial vowel and the second syllable. The dynamic range of the speech spectrum, computed as the difference between the band pressure levels exceeded in 99 and 1% of the measurement intervals, varies from 25.5 dB in the ⅓-octave bands centred at 315 and increases with increasing frequency, to a maximum of 40.5 dB in the 3150 Hz band.

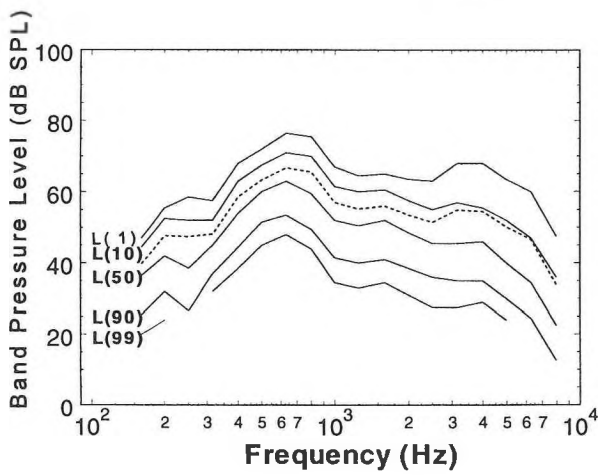


Figure 1. Distribution of the ⅓-octave long-term speech levels for 84 items contained in the UWODFD test. Dashed line is L(eq, 5 min).

To examine the spectrum of the target consonant in isolation from the surrounding context, the target consonants were edited from the test stimuli using a wave-form editor (Jamieson et al., 1992). Formant transitions were included with the consonants. The distribution of these excised consonants is presented in Figure 2. The influence of the adjacent vowels remained visible, however, the dynamic range of the consonant-only portion of the speech materials is

narrower than for the entire nonsense word, particularly in the higher frequency regions.

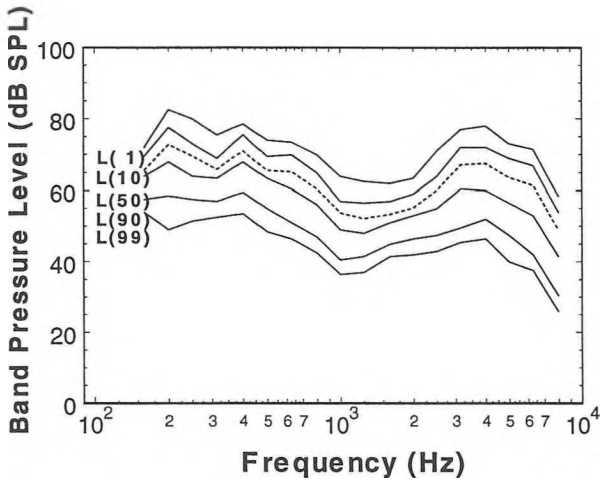


Figure 2. Distribution of the ⅓-octave long-term speech levels of the target consonants only. Dashed line is L(eq, 5 min).

8. BEHAVIOURAL TESTING

8.1. Subjects

Subjects were twenty young adult (age range 20-34 years) staff and students at the University of Western Ontario. All had pure-tone thresholds better than or equal to 20 dB HL (ANSI, 1989) from 250-8000 Hz in the test ear.

8.2. Procedures

No carrier phrase was used during nonsense word presentation. Within the test, stimulus presentation was blocked according to talker and within each talker block, the order of stimulus presentations was randomized without replacement. The listener's task was to choose which consonant was heard from a set of 21 possible responses displayed on a video monitor. The response alternatives were represented on the screen as **b, ch, d, f, g, h, j, k, l, m, n, p, r, s, sh, t, th, v, w, y,** and **z** [1]. Listeners selected one of these response alternatives prior to presentation of the next test item. The complete test of 84 stimuli was used in each speech-in-noise and filtering condition.

8.3.1. Performance-intensity functions. Performance on the test was measured in the presence of a 70 dB(A) noise that was shaped to the ⅓-octave L(eq) of the stimuli (cf. Figure 1 - dashed line). Thirteen signal-to-noise ratios (SNR) ranging from +4 to -20 dB in 2-dB steps were used. Following an initial test in quiet with the speech at 70 dB(A), the test was repeated 13 times, with the order of the SNR for each test randomized for each listener.

Six listeners also completed the test using an audiometer-generated speech-shaped noise masker (Grason Stadtler GSI-16) at eight SNR ranging from -15 to +15 dB and in quiet. The overall speech level was 75 dB SPL.

8.3.2. Filtered speech functions. Fifteen different filtering conditions for the speech stimuli were used: low-pass filtering at 250, 380, 550, 800, 1300, 2300, and 3500 Hz and high-pass filtering at 300, 550, 800, 1300, 2250, 3500, and 5500 Hz and a broadband condition (125 - 8000 Hz). Broadband speech-shaped noise was used in all conditions. The SNR for the equivalent broadband condition was fixed at +4 dB. Following an initial test in the broadband condition, the filtering conditions were completed in a randomized order.

9. RESULTS AND DISCUSSION

9.1. Performance-intensity functions

The mean performance scores as a function of SNR for the broadband listening conditions are shown in Figure 3. The slope of the performance-intensity function is very shallow, averaging 3%/dB in the SNR range from -20 to 0 dB. French and Steinberg (1947) obtained slopes of approximately 5%/dB for their nonsense syllable task and Duggirala, Studebaker, Pavlovic and Sherbecoe (1988) reported slopes of 5.74%/dB for the diagnostic rhyme test. The shallow slope obtained here with the UWODFD may be enhanced by the noise being matched to the combined spectra of the four talkers, rather than to each of the individual talkers (Studebaker, Pavlovic & Sherbecoe, 1987).

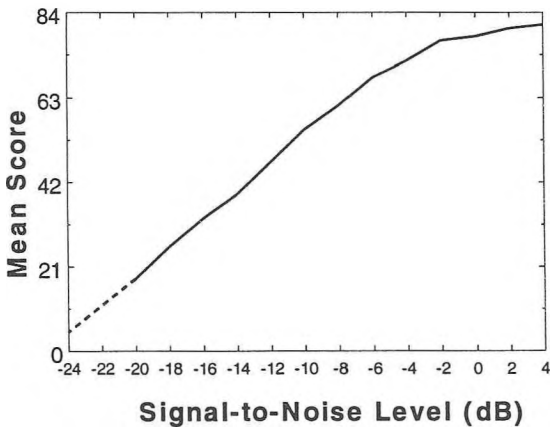


Figure 3. Mean performance scores on the UWODFD test, as a function of the signal to noise level for the broadband listening conditions.

In an independent test with speech shaped noise generated by an GSI-16 audiometer, the mean slope of the linear portion of the performance-intensity functions for 6 subjects, tested from -15 dB SNR to +15dB SNR, was 3.1%/dB. Thus, the very

shallow function for the UWODFD test appears to be a property of the test itself rather than reflecting the specific noise used as a masker

Unlike conversational speech, where higher-level cognitive factors combine with the available acoustic information to produce very steep performance-intensity functions, shallow performance-intensity functions are expected for nonsense syllables. Such a shallow performance-intensity function has a significant advantage for applications where performance differences need to be measured over a wide range of listening conditions.

9.2. Filtered speech functions

The results of the filtered speech conditions are displayed in Figure 4, where the mean score for each of the four blocks (talkers) of the test is shown as a function of cut-off frequency. The crossover frequencies for the high- and low-pass conditions are slightly higher for the female talkers than for the males. The crossover frequency for the test taken as a whole is 2170 Hz, which is higher than that reported by French and Steinberg (1947) for nonsense syllables spoken by male and female talkers, and higher than other reports for nonsense syllables using male voices (Dubno & Dirks, 1989; Duggirala, Studebaker, Pavlovic & Sherbecoe, 1988).

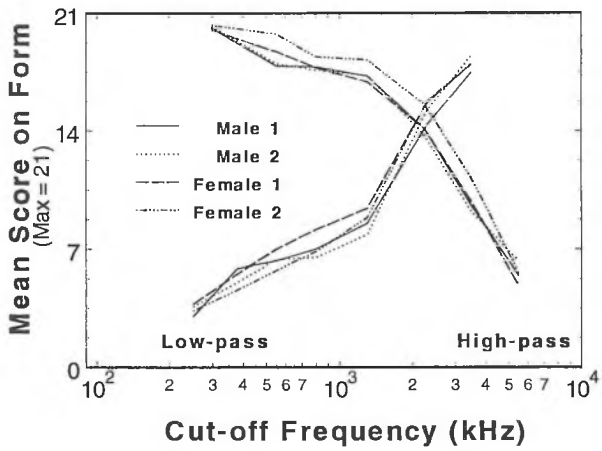


Figure 4. Performance as a function of filter cut-off frequency, for each of four talkers.

9.3. Applicability of conventional Articulation Index weights

Cheesman, Appleyard and Lawrence (1992) reported a series of studies designed to determine whether or not the Articulation Index (ANSI, 1969) frequency-importance weights for nonsense syllables could be applied directly to the UWODFD materials, without modification. ANSI Articulation Index weights did not result in accurate prediction of performance on the UWODFD test, with the fit being particularly poor for the filtering conditions. The dependence of the Articulation Index on a 30-dB dynamic range, which underestimates the dynamic range of the UWODFD materials particularly at higher frequencies (cf., Fig 1), combined with the high cross-over frequency for the UWODFD materials likely contribute to the poor predictive power of the Articulation Index for these materials.

9.4. Comparison of alternative approaches to scoring

The UWODFD test can also be scored using any of a variety of scoring systems based on phonetic feature descriptions of the signals. Feeney and Franks (1982) suggested using a seven-feature scoring system of Voice, Continuant, Strident, High, Back, Anterior, and Coronal for their DFD test. The extension of the stimulus set from 13 consonants to 21 consonants for the UWODFD test required additional feature scoring assignments. The results obtained when the data displayed in Figure 3 are scored using this system, are plotted in Figure 5.

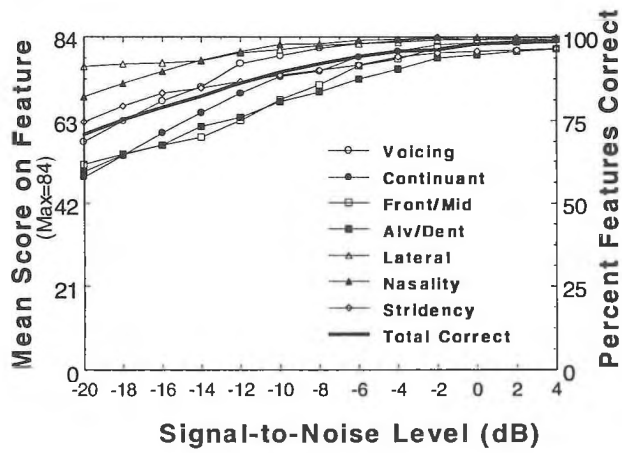


Figure 5. Performance-intensity functions for the data displayed in Figure 3, plotted as the number of correctly-identified features. A separate performance-intensity function is plotted for each of the features analysed (left axis). The total number of correct features (expressed as a percentage) is shown by the solid line (right axis).

Differences in the slope and form of the functions from feature to feature are clear. For example, some features have very low error rates, so they do not contribute to the aggregate curve; for other features, the performance-intensity curve is steeper, indicating that listeners are sensitive to the feature only over a very narrow SNR region.

This seven-feature analysis differs dramatically from the three-feature analysis provided by Cheesman, Lawrence, and Appleyard (1992) as shown in Figure 6. The score for the manner feature is similar to the whole item test score (cf. Figure 3) in the three-feature system. Because the seven-feature system breaks place and manner characteristics into several features each, there are fewer errors on any single place or manner-related features. This results in shallower

performance-intensity functions for both the individual feature functions and for the function of total features correct.

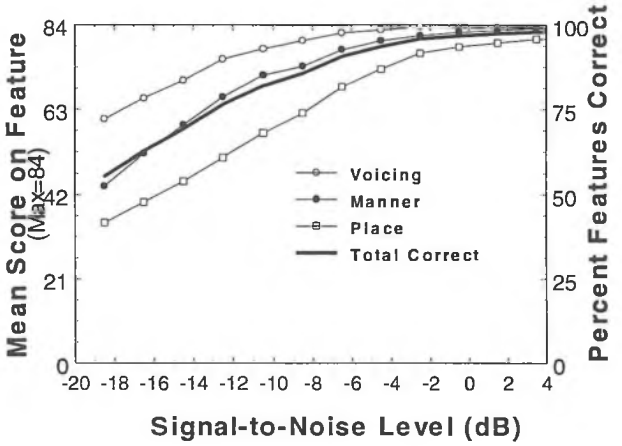


Figure 6. Performance-intensity functions for the three-feature scoring system used by Cheesman, Lawrence, and Appleyard (1992).

9.5. Possible advantages of feature-based scoring

According to Feeney and Franks (1982) and other authors, estimates of subject's performance on a speech intelligibility test such as the DFD are more reliable if data are scored in terms of specific feature errors rather than in terms simply of entire items being correct or incorrect. For example, using whole item scoring, a response of /d/ for /b/ and a response of /t/ for /b/ are equally severe errors. However, in a feature-based scoring approach, the /t/ response is more severe, as /t/ differs from /b/ in both Place of Articulation and Voicing, whereas /d/ differs from /b/, only in Place of Articulation (i.e., Voicing is reported correctly).

Another consideration is that the reliability of the test cannot be predicted readily from the number of test items when items in a test such as the distinctive feature difference test are scored on a feature-by-feature basis. This is because the binomial distribution is unlikely to approximate the test score distribution, because the individual features are not independent and errors are therefore correlated across features. For example, a relatively simple feature scoring system is one in which only place, manner and voicing features are scored as correct or incorrect. If the test item /ʌmɪl/ is presented and the manner feature is correctly perceived (as nasal) then the voicing feature will also be correctly identified, because voiceless nasals are not included in the response set of English consonants.

The data obtained from the filtering conditions can be used to evaluate the proposal that feature-based scoring increases reliability. Reliability coefficients were calculated both when the test was scored on a whole item basis and when the test was scored in terms of the percentage of features correctly identified. The procedures outlined by Winer (1962, **p.** 124) for estimating the reliability of measurements using an analysis of variance model were used.

The estimate of reliability obtained for a single measurement was .52 for both the whole-item and feature scoring approaches. The reliability of the average of the 15 measurements (equivalent to Spearman-Brown reliability, Winer, 1962) was .94 for both scoring methods. Thus, the estimated reliability of the measurements did not differ for the two procedures.

A similar pattern of results obtained with the noise-masked data, for which individual test administration reliability was .37 and .33 for whole-item and feature scoring, respectively, and .89 and .87 for the average of the 14 listening conditions, for whole-item and feature scoring, respectively.

Although these measures do not directly address test-retest reliability under identical listening conditions, they do indicate that, from a test reliability perspective, feature scoring using this seven-feature set does not provide an advantage over the whole-item scoring procedure, despite increasing the number of "scoreable units". This is in contrast to Feeney and Franks' (1982) hypothesis.

Notwithstanding this failure of feature scoring to improve the reliability of speech intelligibility estimates, feature-based scoring may offer an important analytical advantage over traditional speech perception measures. As one example, Jamieson, et al. (1995) used a three-feature (Place, Manner, and Voicing) scoring approach to examine the effects of applying a noise reduction scheme that used a "voicing detector" to toggle the estimate of the background noise provided to the processor. This analysis showed that voicing confusion errors did not increase when the noise reduction scheme was applied. Such a conclusion would not have been possible from consideration of whole-item test results alone and requires the analytic feature approach made possible by the DFD.

10. APPLICATIONS

This modified version of the DFD test has received extensive use in a variety of research projects undertaken by members of Western's Hearing Health Care Research Unit over the past several years. A frequent application has been evaluation of the benefit provided to individual hearing aid users by alternative hearing aid systems. This is a challenging task, requiring high test sensitivity, as the incremental benefit of switching a listener from one carefully-fitted hearing aid to another hearing aid with similar processing characteristics may be relatively small. Jamieson and Cornelisse (1992) used these speech test materials successfully in their evaluation of the differences in listener performance when hearing aid users were fitted with K-amp and linear hearing aids. Jamieson and

Brennan (1992) and Jamieson, Brennan and Cornelisse (1995) used the test successfully to measure the benefit provided to listeners by an adaptive noise reduction filtering system designed for use in future generations of digital hearing aids.

These test materials have also been used as a basic tool to evaluate overall speech intelligibility performance by individual listeners. As one example, Cheesman, Armitage and Marshall (1994) used the UWODFD to measure the speech perception abilities of younger and older Canadians, in a study examining the relation between speech perception ability and growth of masking. Yu and Jamieson (1994) used the UWODFD to quantify changes in the ability of native speakers of the Korean language to identify English-language consonants, following extended exposure to the English language after immigrating to Canada, and throughout the course of a structured program of English-language training.

11. CONCLUSIONS

The studies reviewed here have established that the UWODFD is an appropriate test for a variety of applications requiring measurement of listeners' abilities to identify English language consonants based primarily on acoustic information. The test has been shown to be appropriate for use with subjects from several different educational and cultural backgrounds, it can be administered and scored quickly, it is sensitive and has high reliability. For these reasons, it may prove useful for inclusion as part of a battery of tests for the measurement of spoken language perception. While there is no evidence that feature-based scoring increases the reliability of an overall measure of speech intelligibility performance, such scoring provides a level of analysis not available in conventional approaches.

ACKNOWLEDGEMENTS

This work was funded by the Ontario Ministry of Health, the Ontario Deafness Research Foundation, and NSERC. Valuable assistance was provided by Leonard Cornelisse, Kristina Greenwood, Judy Keith, Lucy Kieffer, Emmet Raftery, Ketan Ramji, and Emmanuelle Ravel. Portions of this report were published in the Proceedings of the International Conference on Spoken Language Processing, Banff, 1992. CSRE and Computerized Speech Research Environment are registered trademarks of Avaaz Innovations Inc.

REFERENCES

American National Standards Institute (1989) *Specifications for audiometers*, ANSI S3.6-1989, New York.

American National Standards Institute (1969) *American National Standard methods for calculation of the Articulation Index*, ANSI S3.5-1969, New York.

Avaaz Innovations Inc. (1995) *CSRE: The Computerized Speech Research Environment.* Software and manual. London, ON.

Boothroyd, A (1968) Statistical theory of the speech discrimination score. *Journal of the Acoustical Society of America, 43,* 362-367.

Cheesman, MF, Lawrence, S and Appleyard, A (1992) Prediction of performance on a nonsense syllable test using the Articulation Index. *Proceedings of the Second International Conference on Spoken Language Processing*, 1123-6.

Cheesman, MF, Armitage, JC and Marshall, K (1994) Speech perception and growth of masking in younger and older adults. *Proceedings of the 1994 International Conference on Spoken Language Processing*, 1951-1954.

Danhauer, JL and Singh, S (1975) *Multidimensional speech perception by the hearing impaired: a treatise on distinctive features* (University Park Press, Baltimore), 1-130.

Dubno, JR and Levitt, H (1981) Predicting consonant confusions from acoustic analysis. *Journal of the Acoustical Society of America, 69,* 249-261.

Dubno, JR and Dirks, DD (1989) Auditory filter characteristics and consonant recognition for hearing-impaired listeners. *Journal of the Acoustical Society of America, 85,* 1666-1675.

Duggirala, V, Studebaker, GA, Pavlovic, CV and Sherbecoe, RL (1988) Frequency importance functions for a feature recognition test material. *Journal of the Acoustical Society of America, 83,* 2372-2382.

Earshen, JJ (1986). Sound measurement: instrumentation and noise descriptors, in B Berger, JC Morrill, WD Ward, and LH Royster (Eds.). *Noise and Hearing Conservation Manual.* Fairfax, VA: American Industrial Hygiene Association.

Feeney, MP and Franks, JR (1982) Test-retest reliability of a distinctive feature difference test for hearing aid evaluation. *Ear and Hearing, 3,* 59-65.

Foster, JR and Haggard, MP (1987) Four alternative auditory feature test (FAAF) -- Linguistic and psychometric properties of the material with normative data in noise. *British Journal of Audiology, 21,* 165-174.

French, NR and Steinberg, JC (1947) Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19,* 90-119.

Gordon-Salant, S (1987) Consonant recognition and confusion patterns among elderly hearing-impaired subjects. *Ear and Hearing, 8,* 270-276.

House, AS, Williams, CE, Hecker, MH, and Kryter, KD (1965) Articulation testing methods: consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37,* 158-166.

Jamieson, DG and Brennan, RL (1992) Evaluation of speech enhancement strategies for normal and hearing impaired listeners. *Proceedings of the European Speech Communication Association Conference on Speech Processing Under*

Adverse Conditions, 3, 1-4.
Jamieson, DG, Brennan, RL, and Cornelisse, LE (1995) Evaluation of a speech enhancement strategy for normal and hearing impaired listeners. *Ear and Hearing, 16,* 274-286.
Jamieson, DG and Cornelisse, LE (1992) Speech processing effects on intelligibility for hearing impaired listeners. *Proceedings of the International Conference on Spoken Language Processing*, Edmonton, University of Alberta, 1035-38.
Jamieson, DG, Ramji, K, Kheirallah, I, and Nearey, T (1992) CSRE: A speech research environment, in JJ Ohala, TM Nearey, BL, Derwing, MM Hodge, and GE Wiebe (Eds). *Proceedings: Second International Conference on Spoken Language Processing,* Edmonton: University of Alberta.
Levitt, H and Resnick, S (1978) Speech reception by the hearing impaired: methods of testing and development of new tests. *Scandinavian Audiology Supplement, 6,* 107-129.
Miller, GA and Nicely, PE (1955) An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 301- 315.
Owens, E and Schubert, ED (1968) The development of consonant items for speech discrimination testing. *Journal of Speech and Hearing Research, 11,* 656-667.
Resnick, SB, Dubno, JR, Hoffnung, S and Levitt, H (1975) Phoneme errors on a nonsense syllable test. *Journal of the Acoustical Society of America, 58,* S114.
Studebaker, GA., Pavlovic, CV and Sherbecoe, RL (1987) A frequency importance function for continuous discourse. *Journal of the Acoustical Society of America, 81,* 1130-1138.
Voiers, WD (1983) Evaluating processed speech using the Diagnostic Rhyme test. *Speech Tech., 1,* 30-39, 1130- 1138.
Winer, BJ (1962) *Statistical principles in experimental design.* New York: McGraw Hill.
Yu, K and Jamieson, DG (1993) Training the English /r/ and /l/ speech contrasts in Korean listeners. *Canadian Acoustics, 21,* 107-108.

NOTES

1. We have since modified our response display screen to provide full orthographic representations of the nonsense words (e.g., abil, achil, adil, afil).

Stimuli are available from the first author at the address listed above or via e-mail at cheesman@uwovax.uwo.ca

ECKEL

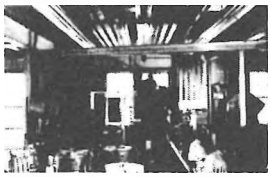
Noise Control Products & Systems

for the protection of personnel... for the proper acoustic environment...

engineered to meet the requirements of Government regulations

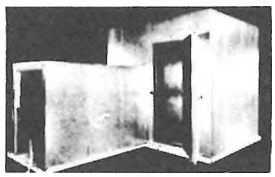
Eckoustic® Functional Panels

Durable, attractive panels having outstanding sound absorption properties. Easy to install. Require little maintenance. EFPs reduce background noise, reverberation, and speech interference; increase efficiency, production, and comfort. Effective sound control in factories, machine shops, computer rooms, laboratories, and wherever people gather to work, play, or relax.



Eckoustic® Enclosures

Modular panels are used to meet numerous acoustic requirements. Typical uses include: machinery enclosures, in-plant offices, partial acoustic enclosures, sound laboratories, production testing areas, environmental test rooms. Eckoustic panels with solid facings on both sides are suitable for constructing reverberation rooms for testing of sound power levels.



Eckoustic® Noise Barrier

● **Noise Reduction Curtain Enclosures** ● **Machinery & Equipment Noise Dampening**

The Eckoustic Noise Barrier provides a unique, efficient method for controlling occupational noise. This Eckoustic sound absorbing-sound attenuating material combination provides excellent noise reduction. The material can be readily mounted on any fixed or movable framework of metal or wood, and used as either a stationary or mobile noise control curtain.

Acoustic Materials & Products for dampening and reducing equipment noise

Multi-Purpose Rooms

Rugged, soundproof enclosures that can be conveniently moved by fork-lift to any area in an industrial or commercial facility. Factory assembled with ventilation and lighting systems. Ideal where a quiet "haven" is desired in a noisy environment: foreman and supervisory offices, Q.C. and product test area, control rooms, construction offices, guard and gate houses, etc.

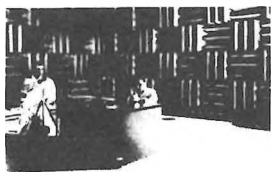
Audiometric Rooms: Survey Booths & Diagnostic Rooms

Eckoustic Audiometric Survey Booths provide proper environment for on-the-spot basic hearing testing. Economical. Portable, with unitized construction.

Diagnostic Rooms offer effective noise reduction for all areas of testing. Designed to meet, within ±3 dB, the requirements of MIL Spec C-81016 (Weps). Nine standard models. Also custom designed facilities.

An-Eck-Oic® Chambers

Echo-free enclosures for acoustic testing and research. Dependable, economical, high performance operation. Both full-size rooms and portable models. Cutoff frequencies up to 300 Hz. Uses include: sound testing of mechanical and electrical machinery, communications equipment, aircraft and automotive equipment, and business machines; noise studies of small electronic equipment, etc.



For more information, contact

ECKEL INDUSTRIES OF CANADA, LTD., Allison Ave., Morrisburg, Ontario • 613-543-2967

ECKEL INDUSTRIES, INC.