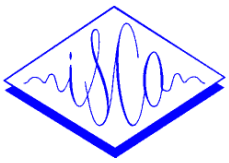

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

PERFORMANCE ON A NONSENSE SYLLABLE TEST USING THE ARTICULATION INDEX

Margaret F. Cheesman, Shelly Lawrence, and Allison Appleyard

Department of Communicative Disorders
University of Western Ontario
London, ON N6G 1H1 CANADA

ABSTRACT

A modification of the Distinctive Features Difference (DFD) test [1] was developed as part of a speech perception test battery, primarily for use in testing new hearing aid circuitry and noise reduction algorithms. The DFD(m) test requires listeners to identify each of 21 consonant sounds, spoken by two male and two female talkers, presented in an /ʌCɑl/ context against a noise background. Identification errors can be scored in terms of either the overall percent-correct item identification or the number of feature differences between each target sound and response. Performance-intensity functions were obtained with a 70 dB(A) background noise masker that was spectrally shaped to match the long-term average spectrum of the 84 test items. Progressive low-pass and high-pass filtering of the speech was used to obtain the crossover frequency of the Articulation Index importance weights for the combined 4-talker speech materials. The crossover frequency was 2170 Hz, higher than that previously found for other sets of nonsense syllables. Articulation Indices were computed using three methods and their relatively suitability for application to the modified DFD test was investigated.

10.21437/ICSLP.1992-332

INTRODUCTION

Analytic consonant perception tests have received increasing attention for use in audiological habilitation, particularly as potential tools for hearing aid evaluation [e.g., 2]. Such tests can be scored in the same way as the tests traditionally used as part of an audiological assessment (e.g., in terms of overall percent correct for W-22 word lists or as the signal-to-noise level required to obtain some fixed level of performance in speech reception threshold testing). However, in addition, the error patterns can be analyzed, either in terms of confusion matrices [3-5] or distinctive feature scoring [1,6].

For distinctive feature scoring, response accuracy is measured in terms of the correctness of the response according to a set of features (e.g., voicing, nasality, continuancy) rather than simply as correct or incorrect identification of the entire consonant. This approach is potentially more sensitive to small differences in listening conditions using a smaller number of test items [1]. The number of feature differences between each target sound and response can be calculated, so that the results of testing with a limited number of items are more stable and reproducible than with whole-item scoring [1]. This efficiency is a particular advantage whenever testing time is limited, as in assessing speech comprehension for hearing aid evaluation, in which more than one speech test or measurement session is likely to occur, or in other clinical applications.

The Distinctive Feature Difference (DFD) Test, developed by Feeney and Franks [1], is a closed-set consonant recognition task in which 13 consonants (/b,t,d,f,dz,k,p,s, ʃ ,t ʃ ,θ,ð,v/) are presented in an /ʌCɑl/ context. The 13 consonants were chosen on the basis of their high likelihood of being erroneously perceived by hearing-impaired listeners in word-initial or word-final positions [7].

In the present study, a modified version of the DFD test was developed DFD(m) that included a larger set of consonants and four different talkers' voices. The larger set of consonants allowed for a wider range of perceptual confusions to occur. The inclusion of more than one talker, although not the norm in speech perception testing, was used to represent speaker variability in the test materials. In order to develop and test the application of the Articulation Index to these materials, the DFD(m) was tested under 14 speech-to-noise levels and under several conditions of high-pass and low-pass filtering using normal-hearing young adult listeners.

METHOD

Stimuli and Instrumentation

The DFD(m) test consists of 84 nonsense words of the form /ʌCɑl/ in which the consonant (C) is one of the 21 consonants /b,t ʃ ,d,f,g,h,j,k,l,m,n,p,r,s, ʃ ,t,ð,v,w,y,z/, spoken by one of four talkers. The talkers were two male and two female young adults from Southern Ontario. During test development, several tokens of each of the words were digitized using the carrier phrase "Point to the word /ʌCɑl/", while ensuring that the peak levels of the phrase stayed constant across tokens. Stimulus sampling to disk was accomplished by low-pass filtering the signal at 8.0 kHz (Kemo VBF 25MD) and 16-bit recording at 20 kHz via an Ariel DSP-16 A/D card. The test tokens were edited from the carrier phrase using CSRE 3.0 software [8]. Each of the final test stimuli was selected from the multiple recordings based on the results of pilot testing using normal-hearing young adult listeners. Items selected were judged to be typical, highly intelligible for normal listeners in quiet and they did not contain idiosyncratic information such as unusual intonation contours or syllable durations that might serve as cues to the identity of the consonant.

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. Stimulus presentation was controlled with a DT-2801A D/A converter and low-pass filtered at 8.0 kHz, except under the filtering conditions specified below. Signal level was controlled using a TTE PA-2 programmable attenuator and an Amcron D-75 amplifier.

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.

Statistical descriptions of the long-term spectrum of the 84 stimuli were obtained through this 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), were measured when the long-term overall level of the speech was adjusted to 70 dB(A).

The statistical distribution of the ⅓-octave long-term speech levels is shown in Figure 1, for the entire DFD test (all 4 talkers). The spectrum is clearly 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.

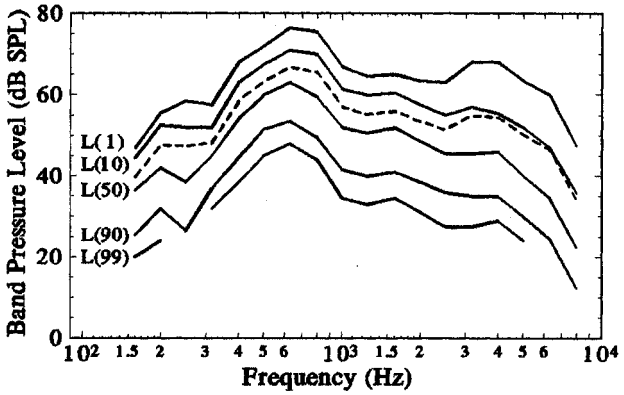

Fig. 1 - Statistical distribution of the long-term ⅓-octave band pressure levels for the DFD(m) materials. The dashed line is the L(eq,5min.).

The ⅓-octave L(eq,5min.) for each of the four talkers are presented in Figure 2. Because of the talker differences and the large contribution of the context surrounding the target consonant to the long-term speech spectrum, there are large differences in the speech spectra for each talker. The pattern of the statistical distribution of the speech levels, as shown for the combined voices in Figure 1, was similar for each talker, with an increase in the dynamic range of the speech spectra at higher frequencies.

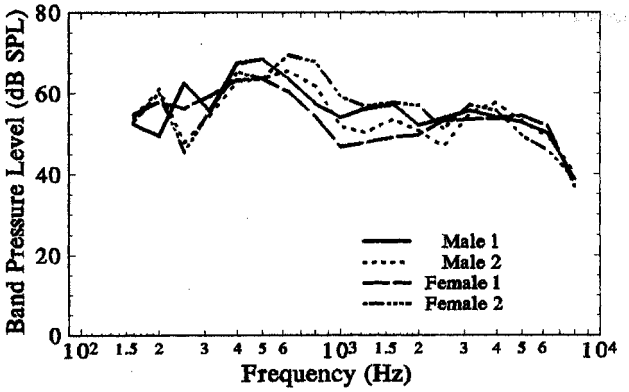

Fig. 2 - Comparison of the long-term ⅓-octave band L(eq,2min.) for each of the 4 voices used in the DFD(m).

The masking noise was generated by a TTE white noise generator and shaped to the ⅓-octave band L(eq) of the 84 stimuli with two Industrial Research Products DG-4017 equalizers applied in series. The full-band long-term L(eq) of the speech-shaped noise was 70 dB(A).

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 or equal to 20 dB HL [9] from 250-8000 Hz in the test ear. All served in each of the performance-intensity and filtered speech experiments.

Procedures

A complete test of the 84 DFD(m) stimuli was run in each speech in noise and filtering condition. Within each test, stimulus presentation was blocked according to talker. The order of the talkers was fixed, with the two male talkers preceding the two female talkers. 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". Listeners were required to select one of these response alternatives prior to presentation of the next stimulus.

Performance-intensity functions. Performance on the test was measured in the presence of the 70 dB(A) speech-shaped noise with 13 signal-to-noise (S/N) levels ranging from +4 to -20 dB in 2-dB steps. Following an initial test in quiet with the speech at 70 dB(A), the test was repeated 13 times, with the order of the S/N levels for each test randomized for each listener.

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. The broadband speech-shaped noise was used in all conditions. The S/N level for the equivalent broadband condition was fixed at +4 dB. Following an initial test in the broadband condition (speech high-pass filtered at 125 Hz and low-pass filtered at 8 kHz), the order of the filtering conditions was randomly selected for each test.

RESULTS AND DISCUSSION

Performance-intensity functions. The mean performance score as a function of S/N level 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 S/N range from -20 to 0 dB. A shallow performance-intensity function for nonsense syllables has been reported frequently [e.g., 11,13]. The shallow slope may be enhanced by the noise being matched to the combined spectra of the four talkers, rather than to each of the individual talkers [10].

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 [11] for nonsense syllables spoken by male and female talkers, and much higher than many other reports for nonsense syllables using male voices [12,13].

Articulation Index. Figure 5 displays the relationship between the overall test score and the AI, when the AI was computed in three ways. In the first two calculations, the ANSI ⅓-octave importance weightings were used [14]. For the first AI computation, labelled "DFD spectrum" in Figure 5, the actual

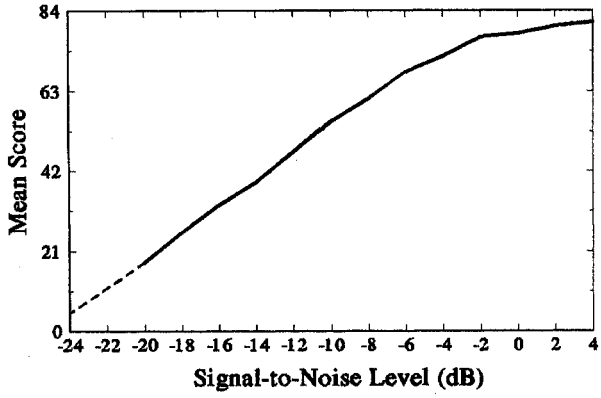

Fig. 3 - The mean DFD(m) score (maximum = 84) as a function of S/N level for 20 listeners.

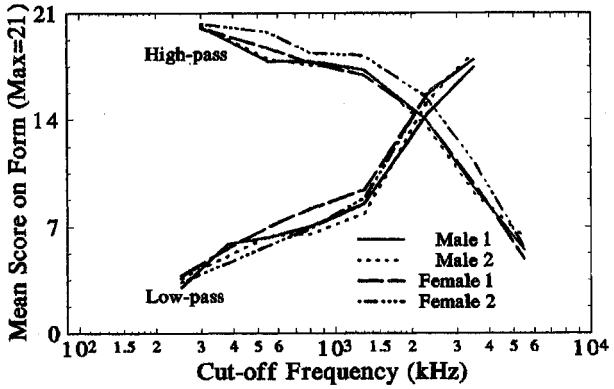

Fig. 4 - The mean DFD(m) score (maximum=21) as a function of filter cut-off frequency for each of the four talkers. Data are averaged across 20 listeners.

speech peaks and noise levels in each ⅓-octave band were used. In the second, labeled "ANSI spectrum", the ANSI speech peaks (adjusted for the overall speech level used in this experiment) and the actual noise levels were used for the AI computation. In both cases, the AI was insensitive to performance differences below approximately 40%, where the computed AI is 0 at speech peak-to-noise levels at which performance is still far better than chance.

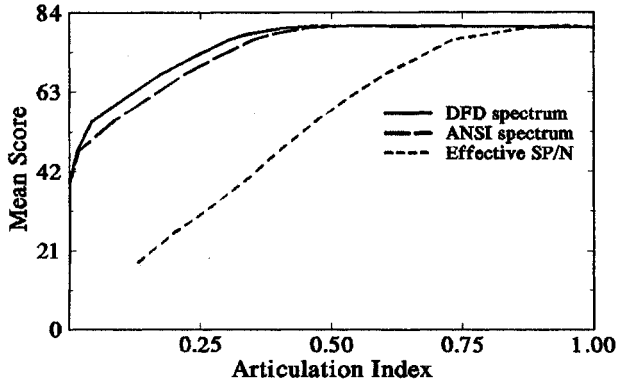

Fig. 5 - The mean DFD(m) score (maximum=84) as a function of AI, where AI calculations are based on 1) the DFD(m) spectra; 2) ANSI spectra; or 3) the effective SP/N.

An alternative method of developing an Articulation Index, and one which deviates somewhat from conventional articulation theory [11] was used by Studebaker, Pavlovic and Sherbecoe [10] during the development of a frequency importance function for continuous discourse materials. Instead of using the absolute S/N ratio to describe the relationship between speech and masker level, the effective speech peak-to-noise ratio (SP/N) was computed. A 0 dB SP/N is "that speech-to-noise relationship that just produces 0% performance" (p.1137) [10] and was obtained for the DFD(m) listening conditions by extending the performance-intensity curve down to chance performance level of 4.76% (dashed line in Figure 3). The S/N level at which chance performance would presumably occur defined 0 SP/N.

Using this performance-based measure of SP/N, the AI was recomputed using the ANSI ⅓-octave importance weights. The dotted line in Figure 5 shows the relationship between the AI computed using the effective SP/N and measured performance.

In order to further examine the relationship between the AI computed using effective SP/N levels and performance, and to further validate the use of the ANSI ⅓-octave importance weights for the DFD(m) materials, the AI was computed for each of the high- and low-pass filtering conditions. The filtered speech test performance as a function of computed AI is plotted in Figure 6. The effective SP/N AI curve is reproduced in this figure as well.

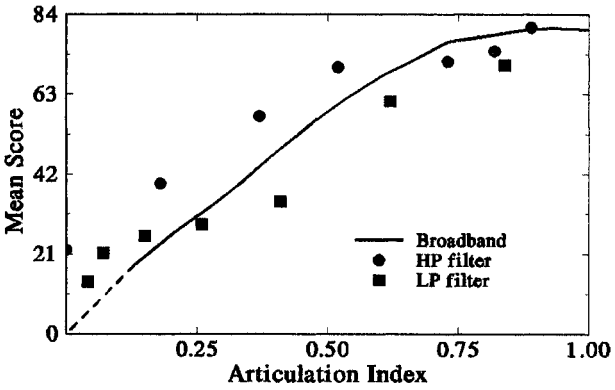

Fig. 6 - Mean DFD(m) performance on each of the filtered speech conditions as a function of the computed AI (solid line). The dashed line is the effective AI curve, replotted from Figure 5, and is based on the broadband listening conditions .

It is clear from this figure that the AI, in the form described by the ANSI standard [14], is not a good predictor of performance on the DFD(m) test. Over a wide range of filtering conditions, higher scores are obtained on the high-pass filtering conditions than for the low-pass conditions that yield comparable AI values. The most probable explanation for this is suggested by the high crossover frequency obtained in the filtered speech conditions; the standard AI weightings underestimate the contribution of the high-frequencies to the identification of the consonants.

Feature Scoring. A preliminary attempt at scoring these materials using a feature-based scheme was made using a single, three feature system. Stimulus-response confusions made in each of the performance-intensity listening conditions were coded in terms of three articulatory features: voicing, place, and manner.

The score for place closely follows the whole item scores, except at poor S/N where manner and voicing errors also contribute to the reduced test score. Duggirala, Studebaker, Pavlovic and Sherbecoe [13] have developed AI frequency importance and transfer functions for each of 6 distinctive features

for the Diagnostic Rhyme Test (DRT). The DRT differs substantially from the DFD(m) test in that it allows for only one feature confusion to occur at a time in a two-alternative forced choice design. Investigation of alternative feature classification or scoring systems, many of which have been reviewed by Danhauer and Singh [6], is currently underway.

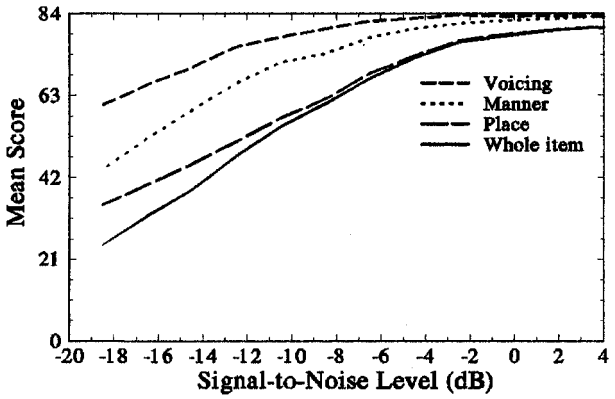

Figure 7. The Mean DFD(m) score (maximum=84) as a function of S/N ratio, using whole item scoring versus feature (voicing, place, manner) scoring. Data are averaged over 20 subjects.

CONCLUSIONS

(1) The crossover frequency for the DFD(m) test is 2170 Hz. The slope of the performance-intensity function is approximately 3%/dB.

(2) ANSI AI procedures [14] do not adequately predict DFD(m) performance under low S/N levels or low-pass or high-pass filtering. Frequency importance weightings and alternative methods for describing the DFD(m) spectra, such as statistical distributions based on shorter intervals than 125 ms or on the information-bearing part of the test items only, are needed before useful AI-type indices can by used for these materials.

(3) In a feature scoring system involving place, manner, and voicing, place errors occurred at high S/N than manner and voicing errors. Voicing errors occurred only at very reduced S/N levels, when overall test scores were below 75%.

REFERENCES

[1] M.P. Feeney and J.R. Franks. "Test-retest reliability of a distinctive feature difference test for hearing aid evaluation," *Ear and Hearing*, 3, pp. 59-65, 1982.

[2] D.G. Jamieson and L. Cornelisse. "Speech processing effects on intelligibility," *2nd International Conference on Spoken Language Processing*, 1992. (in press)

[3] J.R. Dubno and H. Levitt. "Predicting consonant confusions from acoustic analysis," *Journal of the Acoustical Society of America*, 69, pp. 249-261, 1981.

[4] G.A. Miller and P.E. Nicely. "An analysis of perceptual confusions among some English consonants," *Journal of the Acoustical Society of America*, 27, pp. 301-315, 1955.

[5] S. Gordon-Salant. "Consonant recognition and confusion patterns among elderly hearing-impaired subjects," *Ear and Hearing*, 8, pp. 270-276, 1987.

[6] Danhauer, J.L. and S. Singh. 1975. *Multidimensional speech perception by the hearing impaired: a treatise on distinctive features*, University Park Press, Baltimore. 1-130.

[7] E. Owens and E.D. Schubert. "The development of consonant items for speech discrimination testing," *Journal of Speech and Hearing Research*, 11, pp. 656-667, 1968.

[8] D.G. Jamieson, T.M. Neary, and K. Ramji. "CSRE: a speech research environment," *Canadian Acoustics*, 17, pp. 23-35, 1989.

[9] American National Standards Institute, 1969. *Specifications for audiometers*, ANSI S3.6, New York.

[10] G.A. Studebaker, C.V. Pavlovic, and R.L. Sherbecoe. "A frequency importance function for continuous discourse," *Journal of the Acoustical Society of America*, 81, pp. 1130-1138, 1987.

[11] N.R. French and J.C. Steinberg. "Factors governing the intelligibility of speech sounds," *Journal of the Acoustical Society of America*, 19, pp. 90-119, 1947.

[12] J.R. Dubno and D.D. Dirks. "Auditory filter characteristics and consonant recognition for hearing-impaired listeners," *Journal of the Acoustical Society of America*, 85, pp. 1666-1675, 1989.

[13] V. Duggirala, G.A. Studebaker, C.V. Pavlovic, and R.L. Sherbecoe. "Frequency importance functions for a feature recognition test material," *Journal of the Acoustical Society of America*, 83, pp. 2372-2382, 1988.

[14] American National Standards Institute, 1969. *American National Standard methods for calculation of the Articulation Index*, ANSI S3.5, New York.

Acknowledgements

P. Folkeard and K. Greenwood assisted with manuscript preparation. D. Jamieson and D. Fabry provided helpful discussions on the application of the Articulation Index. This work was supported in part by NSERC.