FAAF - AN EFFICIENT ANALYTICAL TEST OF SPEECH PERCEPTION

JOHN R FOSTER AND MARK P HAGGARD

MRC INSTITUTE OF HEARING RESEARCH

The term "speech discrimination score" is often used to describe the results of clinical speech audiometry, though the task is actually an identification task, not a discrimination task, where the listeners would be asked to say whether two things sound the same or different. The usage is tolerated because each identification response generally summates the results of various different types and levels of auditory discrimination process, and this summation underlies the economics of speech audiometry. To avoid fatigue and sequential biasing there is an effective limit of about one identification response every five seconds. The larger the sample of speech material contributing to each such response, the more stable is the test result from a given number of responses but the less circumscribed are the processes which the response can reflect. For specificity as to frequency region or other structural aspects of hearing, long detection or discrimination tests with stimuli have been the norm and appropriate attending and responding skills have needed to be taught. By contrast, natural speech materials exemplify a code in which each response may transmit a large amount of information and may depend upon a large and diverse sample of stimulus information. Minimal training is usually required because of the highly-automated set of identification relationships between stimuli and responses. Hitherto the essential advantages of speech tests have limited the analytical precision they can attain.

The word composition of speech tests has heeded sampling considerations by the use of spondee words, and where a gross index of communication is required, spondee tests or sentence tests have advantages. Spondees also constitute a relatively circumscribed set of about 200 relatively common words. However, more complete control of the effects of verbal ability is possible if we exploit the fact that word-frequency effects disappear in small known vocabularies. Small known vocabularies can also constrain the discrimination to a particular phoneme or feature, the extreme example being a two-choice identification test, which verges on a discrimination test, but has the advantage of maximum ease when the capacity of the source, transmitting channel or receiver happens to be maximally restricted. Where in this continuum of sampling naturalness and analytic power one chooses vocabulary size and other test details depends upon an appraisal of the balance of the various advantages in a particular range of applications.

Objectives and Rationale

We have developed a test to meet a particular set of application constraints:

(1) Ease of response task and lack of practice effects - dictating a small vocabulary printed for each item;

(2) Face validity - precluding synthetic speech at present and precluding use of a fixed small vocabulary throughout;

(3) Omni-regional applicability within the UK and applicability to low literacy levels - precluding vowels as discriminanda;

(4) Sensitivity to mild hearing losses in particular at high frequencies - entailing a preponderance of difficult consonant discriminations.

These considerations led us to the FAAF. The second 'A' is ambiguous but to avoid the commitment of "articulatory" or "acoustic" as we call the test the 4-Alternative Auditory Feature Test. It is a set of principles and options rather than a piece of recorded tape to be bought and sold. It is still capable of further development, eg refinement into several scales, and we give only basic standardisation data and preliminary applications here.

The main principle of the test is the same as that of the 6-alternative Rhyme Test (Fairbanks, 1958), involving sets of minimal pairs of real English words. This, and a previously developed 5-alternative version (Haggard and Mattingly, 1968) did not permit entirely symmetrical minimal pair arrangements. Hence distractor words differing in general features were inevitable, and these were effectively wasted, as they occur very infrequently as false responses. Reducing the vocabulary size to four has largely eliminated this problem. Most subjects and patients can scan a row of four items in five seconds: a vocabulary size of four therefore offers the best compromise between ease, specificity and amount of information per response. The sampling frame for each test word need not be the phoneme inventory of the language, nor need phonemic balance be observed as an overriding principle. We have selected feature distinctions designed to give a particular distribution across the range of difficulty so as to give some resolution of severely limited capacity systems, but to concentrate test power in the region of high capacity (mild loss). The test was not intended to provide a "speech audiogram" ie a graph of the levels at which a certain percentage (eg 50%) of items are heard. That method essentially duplicates diagnostic sensitivity data(the audiogram), although it defines an intensity range of maximum score. Our objective was to measure the identification score in that intensity range with precision, as a function of spectral variables, in order to assess directly the factors generating a maximum.

Method

25 sets of four minimally-paired words were devised giving 100 different words. In the recorded sequence only one of the 4 possible items from each set occurrs on each 25-item page. The response sheets present different orderings of the response set in each of its 4 appearances, coupled with random siting in the column of the actual target word. A program in BASIC for a Cromemco Z2 microcomputer marks and analyses both of two orderings (A & B) of the 100-item test sequence and prints out various indices of the general integrity of the data as well as scores on each of 46 error types.

The error types distinguish the two directions in which a confusion between any two binary feature values can occur, as directional biases frequently occur in hearing loss, and may be sensitive indicators of absence of spectral information. In the set "bail, mail, nail, dale" the error: [voiced labial → voiced dental] could be the result of responding "dale" to "bail", or responding "nail" to "mail". The two reverse errors would be totalled separately. A response of "dale" to "mail" additionally involves nasality and would be an example of a 2-feature error. Such errors are rare and at present do not contribute to the single error scores, which, depending upon the particular feature, are derived from between 2 and 12 possible occurrences. The marking program also produces indices of column response bias and separate overall accuracy scores for initial and final consonants. Our choice of features separate out such distinctions as place of articulation for unvoiced and place of articulation for voiced stop consonants, avoiding the pooling implied by an exclusively articulatory system.

It also at present separates the three binary pairings from the trinary place of articulation feature, and further pooling will be done on an empirical basis only. Apart from totalling across different feature environments for a feature being scored, no attempt is made to force the error patterns into any contrived system such as Distinctive Feature Theory. The particular feature systems derived from multidimensionally scaling consonant confusions in particular levels of noise have not so far proved to be diagnostically useful, and are reflected in the FAAF only in so far as we have biased our items towards difficult distinctions.

Applications

In laboratory experiments on possible speech-processing devices for hearing aids the gross score on the test has been found useful as a sensitive and reliable speech identification measure. In addition, an audiovisual version has been used to measure benefits from a hearing aid in elderly people (Foster, Haggard and Iredale, 1979). In a learning experiment, 6 subjects were presented with Form A, 4 presentations of Form B, then Form A again. The average scores for Form A were 96.2 and 97.2% correct, Form B was mildly distorted for this experiment but gave 90.0, 92.7, 92.2 and 90.0%, confirming lack of practice effects.

In experimental clinical investigations of diminished frequency resolution in cochlear hearing loss due to noise an abbreviated 25-item FAAF in speech-spectrum shaped noise was employed. Correlations were calculated with the 100-item Fry phonetically balanced word list in speech babble and also with estimates of frequency resolution obtained from the psychoacoustical tuning curve for pure tone masking, (Tyler, Fernandes and Wood, 1979). The percentage-correct scores on the FAAF and Fry showed a correlation of r=0.89 over the 10 normal and 12 hearing impaired subjects combined. In a step-wise, multiple regression analysis for the hearing-impaired group, thresholds at the 8 audiometric frequencies explained 87% of FAAF variance as opposed to 81% of the Fry. The correlation for the slope of the psychophysical tuning curve at 4 KHz was 0.65 with the FAAF and 0.63 with the Fry test. Clearly even a 25-item test on the FAAF principle is a reliable clinical instrument.

Standardisation - Stage 1

It is already known that most speech features are spectrally vicarious but that many have a slightly higher weighting in the region of their energy peaks. Binaural presentation of various frequency bands and combinations of bands was designed to determine the frequency-dependence of the various error types. This would enable the simple and robust identification data to furnish a discrimination index for each of 3 or 4 frequency bands, for use in diagnosis as a supplement to the audiogram.

Master recordings of a male and female voice speaking versions A and B were made, and processed by a programmable Finite Impulse Response Filter designed by our colleague Julian Trinder. Bandwidths of 0 to 0.6 KHz, 0 to 1.2 KHz, 0.6 to 1.2 KHz, 1.2 to 2.4 KHz, 0 to 2.4 KHz, 0.6 to 2.4 KHz, 1.2 to 4.8 KHz and 2.4 to 4.8 KHz were used, as well as the two combinations (0 to 0.6)+(1.2 to 2.4) and (0 to 1.2)+(2.4 to 4.8) KHz. Certain dichotic combinations also presented but are not discussed here. The various filter conditions were presented to 108 technical apprentices in a counterbalanced design. The variance over conditions was analysed for each of the 46 error-types. In 43 cases it was significant at the 0.05 level; two of the non-significant types corresponded to the 2 directions of confusion between stop and semivowel categories. This useful exception is explicable in that the distinction is spectrally distributed and based primarily on temporal properties. Predictably, place of articulation distinctions were more difficult than manner or voicing distinctions, the range of error probabilities varying from 0.41 to 0.02 when error data from all filter condiions were combined. Intrusion or omission of the plural morphs /s,z/, of /t/ in clusters, and of velar consonants before vowels were all near the median probability in the distribution, about 0.10. These types of errors are often disregarded in tests, but may yield information on both sensitivity and temporal organisation.

The error data have been used to provide weighted error scores for 4 frequency regions, which give even closer agreement with psychophysical data than do gross scores. They also show that the 46 initially distinguished error types may be reduced by pooling types with closely related empirical distributions, increasing the reliability of each.

Future developments

Standard scores are being obtained for presentation of the bands in noise, and norms will be sought for reference groups of patients. At present there is little reason to change the items in the test sequence as various purposes can be served by focussing on empirically justified subsets of items. A standard protocol for determining rapidly the appropriate levels of presentation for maximum discrimination is under development. The test will be documented and released when these steps are completed.

References

G. FAIRBANKS 1958. Journal of the Acoustical Society of America 30, 596-600. Test of phonemic differentiation: the rhyme test.

J.R. FOSTER, M.P. HAGGARD and F. IREDALE 1979. Submitted for publication. Acoustic, audiological and personal determinants of use of postaural aids in sensory hearing loss.

M.P. HAGGARD 1979 Experimental Brain Research (in Press). Speech sounds in relation to speech processing.

M.P. HAGGARD and I.G. MATTINGLY 1968 IEEE Transactions on Audio and Electroacoustics AU-16, 95-99. A simple program for synthesising British English.

R.S. TYLER, M. FERNANDES and E.J. WOOD 1979. Submitted for publication. Masking, temporal integration and speech intelligibility in individuals with noise-induced hearing-loss.