SCALING OF SPEECH INTELLIGIBILITY USING MAGNITUDE AND CATEGORY ESTIMATION AND PAIRED COMPARISONS

S.C. Purdy and C.V. Pavlovic

University of Iowa, Iowa City, U.S.A. &
University of Provence, France

ABSTRACT

Subjects judged speech intelligibility using either magnitude estimation, category estimation, or paired comparisons. Speech scores for CID Sentences and NU-6 words were also obtained. The speech was bandpass filtered so that a monotonic increase in intelligibility was predicted by articulation index (AI) theory. The reliability and sensitivity of intelligibility judgments and speech scores were compared. The validity of intelligibility judgments was investigated by comparing judged intelligibility to AI predictions.

1. INTRODUCTION

Psychophysical scaling procedures have been used to investigate the intelligibility and quality of hearing-aid transduced speech. These include the methods of paired comparisons (PC) in which subjects judge which of two stimuli has more or less of the attribute being investigated [4, 8], magnitude estimation (ME) in which subjects choose any positive number to represent the subjective magnitude of the attribute present in the stimulus [9, 10], and category estimation (CE) in which subjects rate their impressions by choosing numbers or adjectives from a fixed range of scale values [1, 5]. Subjects can make judgments of speech intelligibility or quality that differentiate reliably among hearing aids. Subjective judgments may be more sensitive to hearing aid differences and more reliable than speech recognition scores [4, 8, 10].

We investigated the reliability, sensitivity, and validity of speech intelligibility judgments obtained using ME, PC, and CE. Validity was tested by comparing intelligibility judgments to AI predictions. According to AI theory, intelligibility is monotonically related to the articulation index: $A = \sum_{i=1}^{n} I_i W_i$, where I_i describes the importance of each frequency band for speech intelligibility, and W_i is a weighting function representing the speech dynamic range contributing to intelligibility [6]. Since the exact form of I_i was not known for our speech, we chose filter settings that produce a monotonic increase in AI for speech materials spanning the probable range of importance functions [6]. Sensitivity was investigated by determining how well procedures differentiated filter conditions. To assess reliability subjects were tested twice

2. PROCEDURE

Thirty subjects aged 60-87 years with hearing better than 30 dB HL at 500-2000 Hz were tested. The 60+ age group was chosen as representative of the majority of hearing aid wearers. Subjects were divided into three groups using ME, CE, and PC, respectively.

The speech was filtered using eight bandpass filter settings: 510-920 Hz, 510-1000 Hz, 510-1100 Hz, 630-1500 Hz, 770-2000 Hz, 700-2000 Hz, 630-2000, and 570-2000 Hz, producing a monotonic increase in the AI for nonsense syllables, easy speech, and average speech [6]. Spectrum density levels were computed for the sentences used for intelligibility judgments, NU-6 words, and CID Sentences. Assuming a +12 to -18 dB dynamic range, the entire signal was audible for all subjects.

Subjects judged the intelligibility of sentences from Grade 8 English texts. A commercial recording of the NU-6 lists was used. Sentences for intelligibility judgments and CID sentences were recorded using an male talker who spoke standard American English. The speech was presented via a TDH-50P earphone. Half the subjects doing ME and half doing CE were given the CID Sentence test. The remaining subjects in these groups were given the NU-6 test. On each visit subjects make 16 practice and 56 test judgments.

For ME subjects assigned a number to match the intelligibility of the sentence. No modulus was used. For CE subjects rated intelligibility using a 20-point scale. Number 1 was marked "very very unintelligible" and number 20 was marked "very very intelligible". For PC each stimulus was paired twice with every other stimulus, with order randomized. Subject decided which sentence was more intelligible. Intelligibility was defined as "..how well you understand the sentence".

3. RESULTS

Visit 1 intelligibility judgments are shown in Figs. 1-3. PC preference scores are the number of times that a filter condition was judged more intelligible than the other in the pair. NU-6 and CID word scores are shown in Figs. 5-6. NU-6 phoneme scores were also analysed.

To compare test-retest reliability Pearson correlation coefficients were calculated between visit 1 and 2 data for each subject. R values (0.51-0.99) were transformed [2] and the means were compared using a Bonferroni-adjusted significance level. Mean test-retest correlations for ME and CE judgments and NU-6 word and phoneme scores did not differ significantly. ME, CE, and NU-6 correlations were higher than correlations for PC and CID Sentences. Test-retest reliability was also investigated by calculating intraclass correlations between visit 1 and 2 data Intraclass correlation coefficients show the absolute similarity between pairs of values. There were no differences in intraclass reliability between ME, CE, and PC. CE and PC judgments and NU-6 scores were significantly more reliable than CID scores.

Relative sensitivity was investigated by

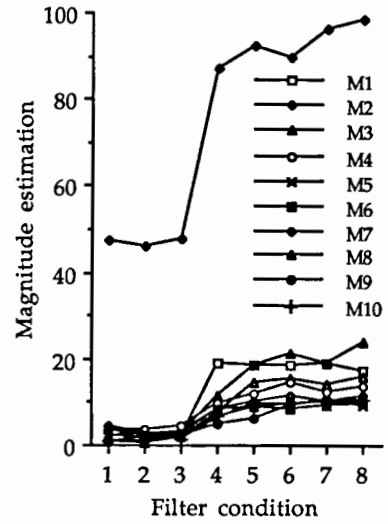

Fig. 1. Individual visit 1 magnitude estimations of speech intelligibility. Each magnitude estimation is the geometric mean of seven judgments.

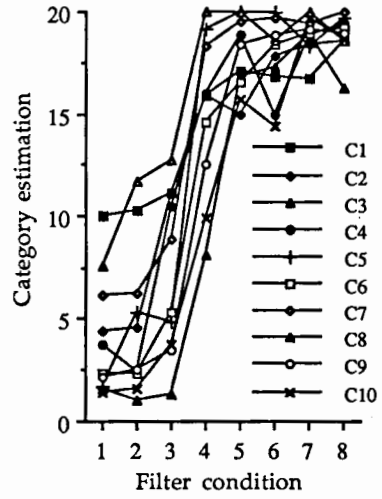

Fig. 2. Individual visit 1 category estimations of speech intelligibility. Each category estimation is the arithmetic mean of seven judgments.

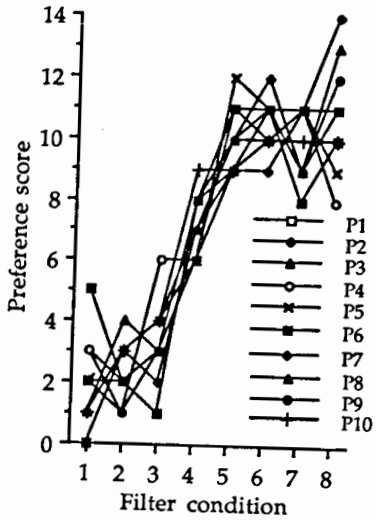

Fig. 3. Individual visit 1 paired comparison judgments of speech intelligibility. Preference scores are the number of times that a filter condition was judged more intelligible.

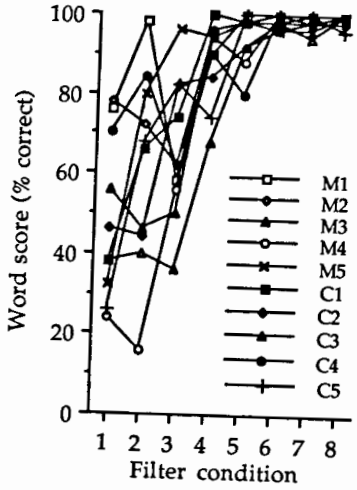

Fig. 4. Individual visit 1 word recognition scores for the CID Everyday Sentence test. Each point is the % of keywords correctly identified in a list.

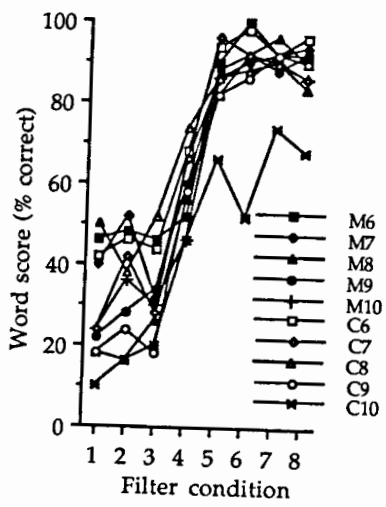

Fig. 5. Individual visit 1 NU-6 word scores. Each point is the % of words correctly identified in a list.

calculating Spearman rank order correlations between intelligibility judgments and AI rankings of intelligibility for the eight filter conditions, and between speech scores and AI ranks. T-tests performed on the transformed r_s values showed no significant differences in sensitivity among the procedures. Because of concern that the lack of statistical power of rank order methods may have disguised any real differences in the data, similar tests of sensitivity were conducted using Pearson's r. Speech scores and intelligibility judgments were correlated with the AI values for nonsense syllables, easy speech, and average speech. For ME, CE, PC, and NU-6 scores, mean correlations were 0.82-0.95. Correlations were generally lower for CID Sentences (0.75-0.86). For all three speech materials there were no significant differences between the correlations for ME, CE, PC, and NU-6 scores. However, intelligibility judgments and NU-6 scores correlated better with the AIs than did CID scores. That is, the CID Sentence test was least sensitive to differences among conditions.

AI theory predicts that, on average, intelligibility should increase monotonically across conditions. Accordingly, if subjective judgments are

valid, intelligibility judgments should be highly correlated with AI rankings of intelligibility for the filter conditions. Even if AI theory is disregarded, an increase in intelligibility is predicted as bandwidth increases across conditions 1 to 3 and 5 to 8. Rank order correlations between the mean intelligibility judgments and the AI rankings were 0.98-1.0. Thus, although individuals confused similar filter conditions, on average there was a monotonic or near-monotonic increase in judged intelligibility across filter conditions, as predicted by AI theory.

4. DISCUSSION

We compared the reliability, sensitivity, and validity of three scaling procedures and two clinical speech recognition tests. Both analyses of test-retest reliability showed that reliability was poorest for CID Sentences. Pearson r values showed that the test-retest reliability of PC judgments was poorer than ME or CE.

There were no differences in sensitivity among the scaling or speech recognition measures based on the analysis of rank order correlations. However, the analysis of Pearson correlation coefficients showed that intelligibility judgments and NU-6 scores were more sensitive to differences between conditions than CID Sentence scores, presumably because CID scores showed the most severe ceiling effect. This is consistent with earlier studies in which speech tests were less sensitive to differences between hearing aids than intelligibility judgments [4, 10].

There were no differences in reliability or sensitivity between NU-6 scores and ME and CE judgments, suggesting that ME or CE could be used instead of speech tests for hearing aid selection. A major advantage of ME and CE procedures over the NU-6 test (and other speech recognition tests) is their efficiency.

The high correlations between mean intelligibility judgments and AI rankings of intelligibility and between mean intelligibility judgments and speech scores suggest that subjects were judging intelligibility and not some other aspect of the speech signal. The good agreement between the data and AI predictions indicate that judgments were primarily based on intelligibility. This is consistent with previous evidence for the validity of intelligibility judgments [1, 3, 5, 7, 8].

5. REFERENCES

[1] COX, R. M., and McDANIEL, D. M. (1984), "Intelligibility ratings of continuous discourse: Application to hearing aid selection", *Journal of the Acoustical Society of America, 76*, 758-766.

[2] FISHER, R. A. (1921), "On the 'probable error' of a coefficient of correlation deduced from a small sample", *Metron, 1*, 1-32.

[3] GRAY, T. F., and SPEAKS, C. E. (1978), "Ability of hearing impaired listeners to understand connected discourse", *Journal of the American Auditory Society, 3*, 159-166.

[4] LEVITT, H., SULLIVAN, J. A., NEUMAN, A. C., and RUBIN-SPITZ, J. A. (1987), "Experiments with a programmable master hearing aid", *Journal of Rehabilitation Research and Development, 24*, 29-54.

[5] NAKATANI, L. H., and DUKES, K. D. (1973), "A sensitive test of speech communication quality", *Journal of the Acoustical Society of America, 53*, 1083-1092.

[6] PAVLOVIC, C. V. (1987), "Derivation of primary parameters and procedures for use in speech intelligibility predictions", *Journal of the Acoustical Society of America, 82*, 413-422.

[7] SPEAKS, C., PARKER, B., HARRIS, C, and KUHL, P. (1972), "Intelligibility of connected discourse", *Journal of Speech and Hearing Research, 15*, 590-602.

[8] STUDEBAKER, G. A., BISSET, J. D., VAN ORT, D. M., and HOFFNUNG, S. (1982), "Paired comparison judgments of relative intelligibility in noise", *Journal of the Acoustical Society of America, 72*, 80-92.

[9] STUDEBAKER, G. A., and SHERBECOE, R. L. (1988), "Magnitude estimations of the intelligibility and quality of speech in noise", *Ear and Hearing, 9*, 259-267.

[10] TECCA, J. E., and GOLDSTEIN, D. P. (1984), "Effect of low-frequency hearing aid response on four measures of speech perception", *Ear and Hearing, 5*, 22-29.

This work has arisen as a result of collaboration between members of the SAM project (Esprit Project 2589) and the University of Iowa.