Intelligibility ratings of continuous discourse: Application to hearing aid selection

Robyn M. Cox and D. Michael McDaniel
Memphis State University, Department of Audiology and Speech Pathology, 807 Jefferson Avenue, Memphis, Tennessee 38105

(Received 24 October 1983; accepted for publication 9 March 1984)

Twelve normal-hearing subjects rated the intelligibility of 35-s, hearing-aid-processed continuous discourse (CD) passages. Three talkers (two male, one female), four hearing aids, and two signal-to-babble (S/B) ratios were used in a completely crossed design. Research questions concerned: (1) ability of listeners to rate intelligibility, (2) sensitivity of hearing aid rankings when rankings were based on intelligibility ratings for three CD passages per instrument, and (3) dependence of hearing aid rankings on (a) S/B ratio, and (b) talker characteristics. Results were: (1) listeners were able to rate intelligibility, (2) rankings based on intelligibility ratings of three CD passages per hearing aid were capable of identifying two superior instruments within a group of four hearing aids that were similar in frequency/gain function, (3) listening in a more difficult S/B ratio substantially decreased the sensitivity of the hearing aid rankings for the female talker but had only minor effects on the rankings for the male talkers, and (4) hearing aid intelligibility rankings were found to be different for different talkers. Applications to hearing aid selection are discussed.

PACS numbers: 43.66.Yw, 43.66.Ts, 43.66.Lj, 43.70.Ep [JH]

INTRODUCTION

Improved perception of everyday speech is usually an important component of the overall benefit provided by a hearing aid. For this reason, a test of speech understanding is often used in an attempt to identify the optimal hearing aid from several instruments that have been preselected for a hearing-impaired individual. However, speech tests employed for hearing aid selection typically lack demonstrated validity in predicting aided benefit in comprehension of everyday conversations. The predictive validity of these speech tests cannot readily be assessed, because no suitable metric for quantifying long-term aided benefit has yet been developed.

In this situation, speech test(s) used for hearing aid selection should be constructed to maximize the *probability* of valid predictions of aided benefit in understanding everyday speech. Because continuous discourse is the type of speech typically encountered by the hearing-impaired individual, the test(s) should incorporate continuous discourse unless another type of test can be shown to be predictive of the client's ability to understand continuous discourse. At this time, there do not appear to be any clinically useful tests that meet this requirement. The most frequently used test (50 item lists of monosyllabic words) has been reported to be a poor predictor of ability to follow continuous discourse (Giolas and Epstein, 1963).

Although listener-produced judgments of the intelligibility of continuous discourse have been utilized informally in hearing aid selection for many years, standardized tests using continuous discourse have been difficult to develop. Giolas (1966) reported a procedure in which a 15-min lecture was scored in terms of the percentage of key words correctly identified in subsequent questions. Such a procedure, although useful in the laboratory, would be unsuitable for clinical use. Adopting a different approach, several investigators have reported on the usefulness of paired comparison judgments of intelligibility of continuous discourse (Zerlin, 1962; Punch and Parker, 1981; Studebaker *et al.*, 1982; and others). This procedure appears to have several advantages including rapid administration, good sensitivity, and high reliability. However, it is difficult to implement validly in a clinical setting because the subject cannot actually wear the hearing aids while he is judging intelligibility—this introduces a variety of possibly confounding factors.

A third approach to quantifying the intelligibility of continuous discourse was investigated by Speaks *et al.* (1972). With this approach, normal-hearing subjects used a percentage scale to rate the intelligibility of 15-s speech-in-noise passages. The results supported the validity of the procedure in that the ratings (1) were a monotonic function of signal-to-noise ratio, and (2) were closely related to measured intelligibility of sentences. Nakatani and Dukes (1973) reported data on a similar measure in which normal-hearing subjects used a scale from 1 to 9 to rate the "understandability" of sentences embedded in competing discourse. Again, the results indicated that the procedure validly quantified intelligibility in that (1) the ratings were monotonically related to the amount of speech degradation produced by additive noise or filtering, and (2) within each of these types of distortion the ratings were sensitive to the degree of degradation.

Gray and Speaks (1978) reported an application of this approach to quantifying intelligibility with a group of hearing-impaired subjects. Their subjects rated the percent intelligibility of 10-s samples of continuous discourse in a voice babble background at three presentation levels and three signal-to-babble (S/B) ratios (hearing aids were not used). Results indicated that mean intelligibility ratings were monotonically related to S/B ratio and to presentation level for all conditions except the 0-dB S/B ratio. In addition, the intrasubject reliability was reported as "reasonable."

The results of these few studies employing intelligibility ratings of continuous discourse are consistent with a proposal that this approach may be suitable for quantifying the intelligibility of hearing-aid-processed continuous discourse in a clinical setting. Potentially, intelligibility ratings could be used for the purpose of selecting, from among several preselected hearing aids, the one that is most likely to provide optimal intelligibility for everyday speech.

One issue that must be addressed in evaluating the usefulness of the intelligibility rating task is the validity of the intelligibility ratings themselves. Because there is no established criterion measure of continuous discourse intelligibility against which the ratings can be compared, it would be necessary to evaluate hypotheses which are directed toward construct rather than criterion validation. This may be accomplished by establishing conditions among which a known rating relationship can be expected *a priori* if subjects are indeed rating intelligibility. For example, conditions which are known to be totally unintelligible should receive very low ratings which do not differ significantly across conditions. Also, conditions which incorporate different amounts of the same degradation, such as background noise, should result in intelligibility ratings that appropriately reflect these differences.

Assuming that intelligibility ratings of hearing-aid-processed speech are found to be valid, it would still be necessary to establish that the sensitivity of the rankings derived from these ratings is adequate to differentiate between hearing aids that may reasonably be expected to be preselected for trial in a hearing aid comparison procedure. Comparative hearing aid evaluations are typically performed using hearing aids that are relatively similar in frequency/gain function. Hence, to be useful in such an evaluation, intelligibility ratings must be capable of differentiating among similar hearing aids.

Finally, if intelligibility ratings are found to be both valid and sensitive enough to differentiate among similar hearing aids, it would be necessary to determine the number of ratings required per hearing aid to generate rankings that are sufficiently reliable to serve as a reasonable basis for a decision in an individual hearing aid evaluation.

The investigation described in this paper was designed to assess the validity and sensitivity of intelligibility ratings when employed in a context similar to that found in hearing aid selections. Subjects provided three intelligibility ratings in each condition. Research questions were as follows:

(1) Will listeners be capable of responding to the intelligibility of hearing-aid-processed speech rather than to unrelated factors? (Intelligibility of speech is defined here as the ability to understand words in the context of meaningful continuous discourse.)

(2) If several hearing aids are ranked on the basis of three intelligibility ratings per instrument, will the rankings be sensitive enough to differentiate between instruments that are fairly similar in frequency/gain function?

(3) If several similar hearing aids are ranked on the basis of three intelligibility ratings per instrument, will the ordering or sensitivity of the rankings vary with signal-to-babble ratio?

(4) If several similar hearing aids are ranked on the basis of three intelligibility ratings per instrument, will the ordering and sensitivity of the rankings be independent of the talker?

I. METHOD

A. Subjects

The subjects were 12 normal hearers (thresholds no poorer than 15 dB HL from 250 through 4000 Hz). Their ages ranged from 23 to 62 with a mean of 33 years. Two were male, 10 were female. The subjects were not college students: most of them were parents of clients at a community speech and hearing center.

Normal-hearing subjects were used, rather than hearing-impaired persons, to avoid confounding the validity and sensitivity of the intelligibility ratings with the often inscrutable effects of hearing losses. Previous studies that have compared normal- and hearing-impaired subjects in terms of the sensitivity of their judgments of hearing-aid-processed speech, have reported very similar results with the two types of listeners (Chial and Daniel, 1977; Punch, 1978; Lawson and Chial, 1982; Studebaker *et al.*, 1982). However, these same studies revealed the intrasubject reliability of the judgments to be somewhat better for the normal hearers than for the hearing-impaired subjects. Hence, clinically applicable information about the reliability of intelligibility ratings could not be derived from this study of normal hearers.

B. Talkers

Three talkers, two males and one female, without discernable regional accents, were selected for the study. One of the male talkers (talker 3) was a professional television broadcaster. Long term rms speech spectra for each talker are shown in Fig. 1. Figure 2 shows a spectrogram of the words "...was first tilled..." spoken by each talker in a mid-sentence context. Although all talkers were highly intelligible to normal hearers, Figs. 1 and 2 reveal differences among them. Figure 1 indicates that the speech of talker 1 (a male) contained less high-frequency energy than that of either talker 2 (a female) or 3 (a male). The spectrograms confirm this impression: talker 1's speech showed little or no energy above the second formant and relatively weak fricatives whereas talkers 2 and 3 consistently produced third and

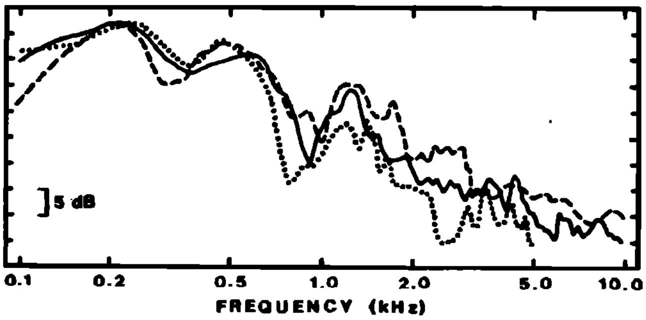

FIG. 1. Long-term rms speech spectrum of each talker. Placement with respect to the ordinate is arbitrary. Dotted line = talker 1, dashed line = talker 2, solid line = talker 3.

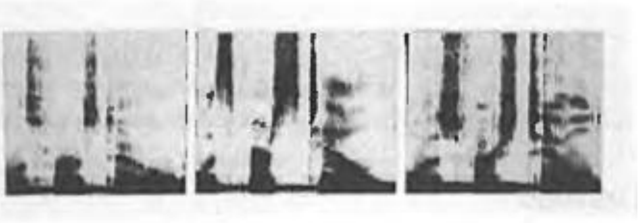

FIG. 2. Spectrogram (0 Hz to 8 kHz) of the words "...was first tilled..." spoken by each talker in a mid-sentence context. Analysis bandwidth was 300 Hz. Talker 1 = left panel, talker 2 = middle panel, talker 3 = right panel.

higher formants and strong fricative energy. In addition, the talkers spoke at somewhat different rates: talkers 1 and 2 both spoke at a rate of approximately 170 words per minute whereas talker 3 spoke more slowly—at about 140 words per minute.

C. Stimuli

The continuous discourse material consisted of 72 passages which were equated *a priori* on the basis of: (1) length—approximately 100 words, requiring 30–40 s to read aloud, (2) subject matter—common plants, animals, and household objects, and (3) vocabulary and sentence structure—the passages were taken from a children's encyclopedia and all conformed to the seventh grade reading level (Fry, 1968). The topic word was contained in the first sentence. Twenty-four passages were randomly allocated to each talker for the production of the stimulus recordings (described below).

The competing stimulus was a multivoice babble recorded in a busy cafeteria. It was edited to remove sections of very high or low intensity.

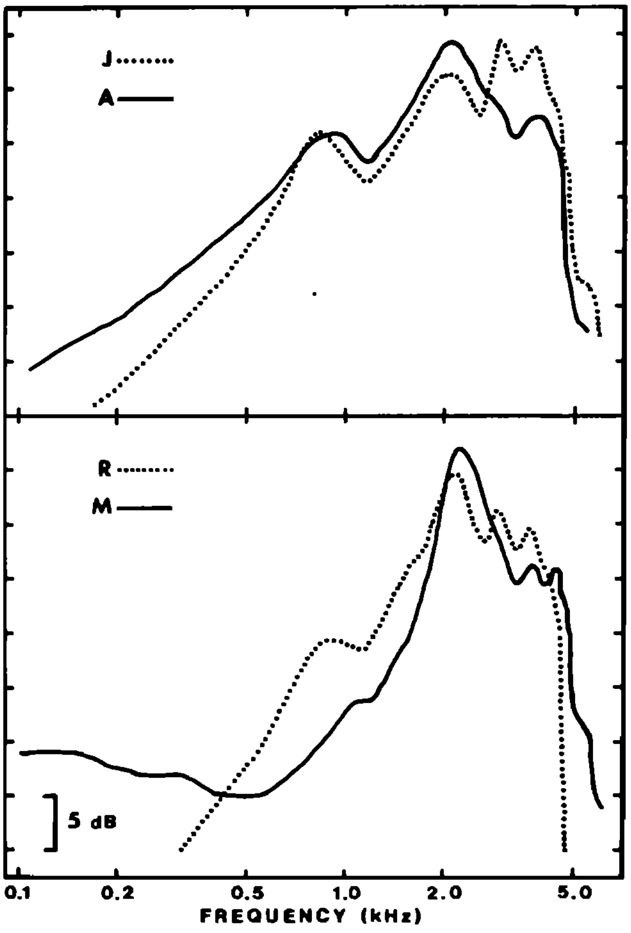

FIG. 3. *In situ* frequency responses of the four hearing aids. Placement with respect to the ordinate is arbitrary. Volume controls were adjusted to equate the outputs in terms of loudness. See Table I for HA-2 coupler gain data.

D. Hearing aids

Four post-auricular hearing aids were used. *In situ* frequency responses for the hearing aids are shown in Fig. 3. These data were measured at the eardrum position of a KEMAR manikin that wore the hearing aids in a sound-treated audiometric room. The curves were obtained using a 400-line spectrum analyzer (Wavetek Rockland, model 5820A), set to the 10-kHz bandwidth. The input was a white noise, equalized to produce a flat spectrum in the sound field, presented from a 0° azimuth loudspeaker at 65 dB SPL. Table I shows the HA-2 coupler gain at the used volume setting for each of the four hearing aids (identified as J, A, R, and M). The four instruments were rather similar in high-frequency average gain: 24, 28, 28, and 27 dB, respectively, although they differed in the details of their frequency responses. These hearing aids were selected because they were sufficiently similar to each other that they might reasonably be preselected for comparison in a hearing aid selection procedure.

The four hearing aids had been used in previous research (Studebaker *et al.*, 1982) and were those identified as J, A, R, and M, in that report. Mean scores obtained by normal-hearing subjects on NU #6 monosyllabic words processed through these instruments at 0-dB S/B ratio were known (74.7%, 71.4%, 64.1%, and 52.5%, respectively). Although these data were interesting for comparative purposes, they were not intended to be used as a validation criterion.

E. Production of stimulus recordings

1. Original recordings

Twenty-four different continuous discourse (CD) passages were tape recorded by each talker. Peak VU readings indicated that all of the passages for a given talker were equal in level within a range of plus or minus 0.5 dB. These passages were subsequently hearing-aid-processed to produce the test tapes. In addition, each talker recorded a 1-min passage on the topic of scarecrows. This passage was subse-

TABLE I. HA-2 coupler gain (dB) at standard test frequencies for the four hearing aids used in this study (identified as J, A, R, and M). The volume controls were set as used for production of the stimulus recordings.

Hz	J	A	R	M
500	15	30	13	−2
800	30	31	24	12
1000	27	30	28	23
1600	25	25	27	24
2500	21	29	28	35
4000	19	17	21	22

quently used to familiarize the subjects with the different S/B ratios and the speaking characteristics of the different talkers. Finally, for each talker, 30 s of the scarecrow passage were recorded backwards which produced a natural sounding but unintelligible result. These unintelligible passages were also hearing-aid-processed for use in "catch" trials to determine whether subjects were actually basing their judgments on intelligibility.

2. Hearing-aid-processed recordings

The recordings made by each talker were mixed with the speech babble and delivered at 65 dB SPL by a wall-mounted loudspeaker into an audiometric test room. The hearing aids were worn by a KEMAR manikin located 1 m in front of the loudspeaker. They were coupled using a full earmold incorporating #13 tubing terminating in a 15-mm section of 4-mm bore. The gain control settings were adjusted so that the four hearing aids' outputs *in situ* were equated on the basis of loudness (Zwicker method, ISO Recommendation R 532, December 1966). The resulting HA-2 coupler gain values are shown in Table I.

A pilot study was performed to select two S/B ratios that were labeled "easy" and "hard." The easy S/B ratio was intended to be fairly, but not completely, intelligible. The hard S/B ratio was expected to be difficult, but not impossible, to understand. For talkers 2 and 3 the selected S/B ratios were + 7 dB (easy) and + 4 dB (hard). Talker 1 was found to be inherently less intelligible than 2 or 3 and was therefore recorded at S/B ratios of + 12 dB (easy) and + 9 dB (hard).

For each talker, 12 CD passages were hearing-aid-processed at each S/B ratio: three CD passages through each of the four hearing aids. In addition, catch passages at each S/B ratio were recorded through three randomly selected hearing aids for each talker.

F. Procedures

The instrumentation used to present the prerecorded test stimuli to the subjects has been described in detail elsewhere (Cox and Studebaker, 1980). Briefly, the stimuli were played on one channel of a tape recorder (Revox A77), delivered to an amplifier, a locally made equalizer, a dc bias device, and a Knowles BP 1712 receiver. The receiver delivered the acoustic stimuli to a length of damped tubing sealed into a compressible foam earplug that simulated a standard earmold. The frequency response of the playback system was flat, plus or minus 2.5 dB, from 100 through 6200 Hz. All testing was monaural: the nontest ear was plugged.

To control sequencing effects, one of the three passages recorded for each hearing aid condition was randomly assigned to each of three separate test sessions. Each session incorporated a CD passage processed through each hearing aid and one hearing-aid-processed catch passage at each S/B ratio for each talker. No CD passage was ever repeated. Within each session, all experimental variables were counterbalanced or randomized. Sessions were separated by at least one day, usually by several days.

Subjects rated the intelligibility of hearing-aid-processed passages on an equal appearing interval scale from 0 to 10. For each talker at each S/B ratio, the sequence was as described in the following excerpt from the subject's instructions:

"... First you will hear one of the talkers reading a passage about scarecrows which is about 1-min long. This is to allow you to become familiar with his or her voice and the amount of background noise. You will not need to do anything at this time except listen and try to understand what is being said.

Next, you will hear five shorter passages by the same talker. At the end of each passage the tape will be stopped and you will be given time to mark on the sheets of paper provided how well you understood the words spoken. You will use a scale of 0 to 10.

If you mark 0 it means that you understood *none* of the words; 10 means that you understood *all* of the words. You should use the numbers between 1 and 9 if you understood some of the words but not all of them. For example, if you think you understood about half of the words, you should give that passage a score of 5. If you only missed a few words, give the passage a 9. On the other hand, if you only understood a few words, you should give the passage about a 1. If you understood about 30% of the words, give the passage a 3, and so on. Since each passage has about 100 words, you could move up one number in score for every 10 words you understood... ."

The "five shorter passages" mentioned in these instructions included four hearing-aid-processed CD passages and a catch passage. In each test session, this sequence of one familiarization passage followed by five test passages was performed for all talkers at both S/B ratios.

The level at which the stimuli were delivered was empirically determined with the aim of avoiding ratings of 10 and/or 0. This resulted in a presentation level (based on peak VU readings) of 37 dB SPL for talkers 2 and 3 and 55 dB SPL for talker 1. For a given talker, the presentation level was the same for all subjects.

TABLE II. Summary of significant ($p < 0.05$) main effects and interactions in four-way analysis of variance of rating data.

Source	df	ms	F	prob
Talker (A)	2	231.8	25.8	<0.001
Error	22	9.0		
S/B ratio (B)	1	1851.3	229.5	<0.001
Error	11	8.1		
A×B	2	174.4	82.6	<0.001
Error	22	2.1		
H. aid cond. (D)	4	1124.4	101.9	<0.001
Error	44	11.0		
A×D	8	47.4	14.3	<0.001
Error	88	3.3		
B×D	4	137.7	59.2	<0.001
Error	44	2.3		
A×B×D	8	12.6	6.3	<0.001
Error	88	2.0		

II. RESULTS AND DISCUSSION

Ratings assigned to the various listening conditions were subjected to a four-factor analysis of variance (3×2×3×5) with the following variables: talkers, S/B ratios, test sessions, hearing aid conditions (four hearing aids and the catch passage) (Winer, 1971). The significant results are summarized in Table II. The test sessions variable did not result in any significant effects. Hence, best estimates for both ratings and rankings were determined by combining the data for the three CD passages delivered in each condition.

Hypotheses concerning the validity of the intelligibility ratings were evaluated using the rating data, i.e., the actual numbers assigned by subjects to the CD passages in the various listening conditions. Issues which were directly relevant to hearing aid selections (sensitivity of rankings, effects of S/B ratio and talkers) were evaluated using ranking data because in this context the ranking of the instruments is the salient issue: one is essentially interested in identifying the best hearing aid(s) from the preselected group.

A. Validity of intelligibility ratings

To evaluate whether the subjects were actually responding to the intelligibility of the CD passages or to some other feature of these stimuli, it was necessary to utilize hypotheses derived from considerations of construct validity. For example, although it was not obvious what the intelligibility relationship should have been among the four hearing aids, it was clear that the hearing aids should have received higher intelligibility ratings than the catch passages. Also, it was apparent that the easy S/B ratio conditions should have been assigned higher ratings than the analogous hard S/B ratio conditions. Furthermore, the various catch passages would be expected to receive ratings which were very low and not significantly different from each other.

These matters were investigated using the analysis of variance of the rating data summarized in Table II. Relevant interactions were investigated using a least significant differences modified *post hoc* analysis (Winer, 1971). Table III shows the results of the *post hoc* analysis of the talker×S/B ratio×hearing aid conditions interaction. This interaction was significant in the main analysis of variance ($p<0.001$). The hearing aids are identified as J, A, R, and M. The catch passage is identified as C. Mean intelligibility ratings are given. Within each talker–S/B ratio combination, the hearing aid conditions are ordered according to the mean ratings. Conditions for which the ratings were not significantly different ($p<0.01$) are underlined. Consideration of these data reveals the following: (1) in the easy S/B ratio the four hearing aids were rated significantly more intelligible than the catch passage for all three talkers. Also, the catch passage received extremely low mean intelligibility ratings—consistent with the fact that it was actually unintelligible. (2) In the hard S/B ratio this situation changed to the extent that hearing aid M was not significantly differentiated from the catch passage for talkers 2 and 3. Because the mean ratings for hearing aid M in the hard S/B ratio were very low for talkers 2 and 3, it would appear that in this condition the speech processed by hearing aid M was itself almost unintelligible.

Comparison of mean ratings across S/B ratios indicates that for a given hearing aid–talker combination the hard S/B ratio always produced the lower intelligibility rating. Analysis revealed that these differences were all significant ($p<0.001$). However, ratings for the catch passages were not significantly different for the two S/B ratios for any talker or for the three talkers in either S/B ratio.

These results support the several hypotheses noted above. This outcome indicates that the subjects were, as instructed, rating the intelligibility of the stimuli.

B. Sensitivity of rankings based on intelligibility ratings

Because the four hearing aids used in the present study were all rather similar in frequency/gain function, it was possible that the intelligibility differences between them were too small to be resolved by the intelligibility rating procedure. In other words, the fact that the hearing aids were ranked in a particular order under certain conditions may have been a matter of chance rather than due to significant differences among them. As discussed earlier, intelligibility ratings could not be usefully employed in hearing aid selections unless the rankings derived from these ratings are capable of differentiating significantly among similar hearing aids.

TABLE III. Results of *post hoc* analysis of rating data for the talker × S/B ratio × hearing aid conditions interaction. The hearing aids are identified as J, A, R, and M. The hearing-aid-processed catch trial is identified as C. Mean intelligibility rating for three CD passages is given for each hearing aid condition. The five hearing aid conditions are ordered according to these mean ratings for each talker–S/B ratio. Underlining indicates the conditions for which the ratings were not significantly different ($p<0.01$). HAC = hearing aid condition.

Talker		Easy S/B ratio					Hard S/B ratio				
1	HAC =	A	R	J	M	C	A	R	J	M	C
	Rating =	9.3	7.9	6.6	5.1	0.5	7.1	5.6	5.4	4.1	0.8
2	HAC =	A	J	R	M	C	A	R	J	M	C
	Rating =	8.5	8.0	7.5	6.2	0.7	2.9	2.6	2.3	2.1	0.4
3	HAC =	J	R	A	M	C	J	R	A	M	C
	Rating =	7.8	6.6	6.4	3.1	0.7	3.9	3.1	2.6	1.9	0.4

TABLE IV. Results of *post hoc* analyses of Friedman's two-way analysis (Siegel, 1956) of variance by ranks for each talker–S/B ratio combination. The hearing aids are identified as J, A, and R. For each subject, the three hearing aids were ranked on the basis of the combined intelligibility ratings for three CD passages per instrument. The sum of ranks for each hearing aid across the 12 subjects is given and the hearing aids are ordered according to these sums. Underlining indicates the conditions for which the ranks were not significantly different ($p < 0.05$). HA = hearing aid.

Talker		Easy S/B ratio			Hard S/B ratio		
1	HA =	A	R	J	A	J	R
	Summed ranks =	35.5	23.5	13.0	31.0	23.0	18.0
2	HA =	A	J	R	R	A	J
	Summed ranks =	30.5	24.5	17.0	26.0	25.0	21.0
3	HA =	J	R	A	J	R	A
	Summed ranks =	31.0	21.5	19.5	33.0	23.0	16.0

To investigate this issue, the rank data for hearing aids J, A, R, and M were subjected to Friedman two-way analyses of variance by ranks (Siegel, 1956). Because the parametric analysis of variance summarized in Table II revealed a significant talker × S/B ratio × hearing aid condition interaction, six separate analyses were performed on the rank data: one for each talker–S/B ratio combination. In each analysis, the hearing aid ranks for each subject were determined by: (1) summing the ratings for the three CD passages presented for each hearing aid, and (2) assigning ranks 1 through 4 to J, A, R, and M on the basis of the total rating score for each instrument, beginning with the lowest score. Each analysis revealed the probability that the rank orders assigned to all four hearing aids by the 12 subjects were determined by chance.

The results of five of the six analyses were significant ($p < 0.001$), indicating that at least two of the four hearing aids were ranked significantly differently. In the sixth analysis—talker 2 in the hard S/B ratio—the result was not significant, indicating that none of the hearing aids was significantly differentiated from the others. In the five significant analyses, inspection of the data suggested that the significant result may have been due mainly to a strong tendency for hearing aid M to receive the lowest ranking (rank no. 1). *Post hoc* comparisons using Nemenyi's test (Kirk, 1968) confirmed that hearing aid M was always ranked significantly lower than at least one other instrument and usually lower than two or more.

Certain independent evidence was available to suggest that hearing aids J, A, and R would be rather similar to each other in terms of intelligibility for continuous discourse and better than hearing aid M. First, the frequency response bandwidths of J, A, and R were all wider than that of M and quite similar to each other (see Fig. 3). Second, as mentioned earlier, the mean monosyllabic word intelligibility scores for J, A, and R were within a range of plus or minus 5% whereas the score for M was considerably poorer (Studebaker *et al.*, 1982). The finding that hearing aid M was the lowest ranked instrument overall was consistent with these observations and indicated that the intelligibility rankings were capable of rejecting this apparently inferior instrument. In addition, it was of particular interest to assess the ability of the speech intelligibility ratings to produce rankings which differentiated between hearing aids J, A, and R.

To evaluate this issue, the six Friedman's analyses of variance by ranks were recomputed using only hearing aids J, A, and R. These hearing aids were ranked from 1 to 3 in order of increasing rated intelligibility. Hearing aid M was omitted from these analyses because its consistently lowest ranking obscured the issue in question of whether J, A, and R were significantly differentiated from each other. The results indicated that the same five analyses were again significant ($p < 0.05$). Talker 2 in the hard S/B ratio was the only condition which did not result in significant differentiation between at least two of these three hearing aids. *Post hoc* comparisons were performed using Nemenyi's test (Kirk, 1968). The results are shown in Table IV. As Table IV shows, the typical outcome was for the no. 3 ranked (best) hearing aid to be significantly differentiated from the no. 1 ranked instrument but not from the no. 2 ranked instrument. This is less than an ideal outcome from the point of view of application of this procedure to hearing aid selection: in the ideal case, the best hearing aid would be significantly differentiated from all of the other instruments. Only talker 1 in the easy S/B ratio achieved this ideal. However, the ability to validly identify the two best instruments from a group of four similar preselected hearing aids would be of considerable benefit in many instances.

C. Effect of signal-to-babble ratio

To differentiate among hearing aids on the basis of intelligibility in a clinical setting, it is very often necessary to introduce a competing stimulus in order to prevent some or all of the conditions from being fully intelligible. It is not usual to standardize the S/B ratio for this purpose since the appropriate value varies with the abilities of the hearing-impaired person as well as with such issues as the type of competing stimulus, the inherent intelligibility of the talker, the semantic and phonemic content of the speech, etc. As a result, it is usually necessary to select the S/B ratio during the clinical evaluation process and this is often done in a rather unsystematic way. It is relevant, therefore, to consider whether the S/B ratio itself affected the hearing aid rankings or the ability of the subjects to differentiate between the instruments.

Table IV reveals the effects of the change in S/B ratio on the rankings for J, A, and R. It should be recalled that hearing aid M received the lowest ranking in all six conditions. The effects on the overall hearing aid rankings of changing the S/B ratio varied across the talkers. The rankings for talker 3 were independent of S/B ratio. For talker 2, the same instrument was ranked best in both S/B ratios but the instruments ranked no. 1 and no. 2 were transposed. For talker 3, all three hearing aids assumed different overall rankings in the hard S/B ratio but this observation is not very meaningful since the rankings for talker 3 in the hard S/B ratio have been shown to be a matter of chance.

As suggested in the previous sentence, there was some indication that the speech intelligibility rankings may have been less sensitive when employed under more difficult listening conditions. As Table IV shows, in the easy S/B ratio the rankings for at least two of the three hearing aids were significantly differentiated for all talkers. However, in the hard S/B ratio, the rankings for talker 2 were not significantly different for the three instruments. Also, for talker 1, a significant distinction between hearing aids A and J in the easy S/B ratio was not maintained in the hard S/B ratio condition.

These data suggest that the sensitivity of the speech intelligibility rankings is adequate to provide useful, though not perfect, differentiation among similar hearing aids as long as the listening conditions are not too difficult. In more difficult listening circumstances, however, the sensitivity of the rankings seems to depend more heavily on the characteristics of the talker. The results further suggest that if adequate sensitivity is maintained for a particular talker, the same hearing aid would be ranked most intelligible in both easy and hard S/B ratios.

D. Effect of talker characteristics

It is conceivable that hearing aid–talker interactions occur on the dimension of speech intelligibility. As an example, if a talker's voice contains mostly low-frequency energy, the different abilities of hearing aids to reproduce high frequencies may be irrelevant to that talker's intelligibility. On the other hand, if the talker's voice contains a large proportion of high-frequency content, different high-frequency capabilities in hearing aids might be expected to influence intelligibility for that talker's speech. Furthermore, characteristics of speech other than its spectrum might interact with the intelligibility of hearing-aid-processed speech. For instance, speech rate could combine with transient distortion in hearing aids with the result that for a relatively rapid talker, differences in transient distortion may be decisive determinants of intelligibility whereas in a slower talker they may be less important.

In the present investigation, the effects of the talker were studied at a relatively superficial level. The research question was simply: Do hearing aid intelligibility rankings interact with talker characteristics? The results for hearing aids J, A, and R may be seen in Table IV. Again, it should be recalled that hearing aid M received the lowest ranking in all conditions. Hence, the ranking of this clearly inferior instrument was independent of the talker.

In the easy S/B ratio, each talker produced a different ranking of hearing aids A, R, and J. The Friedman's analyses of variance described earlier (Siegel, 1956) revealed that the rank order produced for each talker was statistically significant ($p < 0.05$), even though the ordering of the hearing aids was different for each talker. Although different, the ranks produced for talkers 1 and 2 had much in common: both had hearing aid A ranked best and significantly superior to hearing aid R with no difference between hearing aids R and J. However, talker 3 produced an entirely different result with hearing aid J ranked best and significantly superior to hearing aid A. These data are consistent with a hypothesis of a hearing aid rank × talker interaction.

Table IV also shows, as remarked earlier, that the ranks for talker 3 remained unchanged across both S/B ratios and relatively minor changes were seen in talker 1's rankings across these conditions. On the other hand, for talker 2 in the hard S/B ratio, the rankings of the three hearing aids were a matter of chance. These results suggest not only that hearing aid rankings interact with talker characteristics, but also that certain talkers may produce more robust results—i.e., results that are obtained over a wider range of test conditions. In the present instance, talker 3 produced more robust rankings than talker 1 who, in turn, performed better than talker 2.

Since talkers 1 and 3 were male and talker 2 was female, the possibility arises that these intertalker differences were at least partially related to intrinsic differences between male and female voices. However, this suggestion must be made tentatively in the present study since there were some differences in the conditions under which the three talkers were tested: talker 1 was presented at a higher listening level and a different S/B ratio than talkers 2 and 3. The effect, if any, of these differences is not known.

When talker is held constant, intelligibility of monosyllables has been found to vary with amplification system bandwidth and S/B ratio (Skinner and Miller, 1983). Overall bandwidth and S/B ratio undoubtedly affect the intelligibility of hearing-aid-processed continuous discourse also. Since individual talkers are known to vary considerably in factors such as speech spectrum bandwidth, interharmonic spacing, and speaking rate, it should not be surprising that the results of this investigation suggest that talker characteristics may interact with hearing aid characteristics to determine which of a group of similar hearing aids will be ranked most intelligible when the hearing aids are processing continuous discourse.

III. APPLICATION TO HEARING AID SELECTION

The results of this investigation using normal hearers indicated that the intelligibility rating task can result in valid quantification of the intelligibility of hearing-aid-processed continuous discourse. Furthermore, as long as the listening conditions were fairly (but not completely) intelligible, all three talkers participating in this study produced hearing aid rankings which significantly differentiated among the four similar hearing aids used. Although a single unequivocally "best" hearing aid was not often identified by the intelligibility rankings, at least two of the four instruments were shown to be inferior in intelligibility. In most comparative hearing aid evaluations, the narrowing of selection alternatives thus achieved would be of considerable benefit. In work with hearing-impaired persons, sensitivity would be maximized by individual selection of S/B ratio for each listener/talker combination. The aim of this adjustment would be to avoid ratings of 10 while preventing the S/B ratio from becoming so difficult that the effect is to obscure differences among hearing aids.

The ranking of hearing aids in terms of the intelligibility they provide when processing continuous discourse was not found to be independent of the talker. This outcome has important implications for comparative hearing aid evaluation procedures since it indicates that the hearing aid recommended may be determined, in part, by the talker used in the evaluation. It would appear necessary, therefore, to define the appropriate spectral and/or temporal characteristics for talkers whose speech is used to assist in hearing aid selection. Several possibilities exist. Perhaps talker characteristics should be chosen: (1) to produce the greatest distinctions between hearing aids, (2) to result in the best possible intelligibility, (3) to be equal to the average characteristics of the adult population, or (4) to mimic the characteristics of the most significant talker(s) in the everyday life of the hearing-impaired individual. If digital techniques were used for storage and manipulation of speech, it would be feasible to select talker characteristics to suit the needs of the client.

In addition to further study of intertalker differences, future work in applying the intelligibility rating task to hearing aid selection will employ hearing-impaired subjects to determine the number of different ratings typically required per hearing aid to produce reliable rankings. In the present study with normal hearers, the rankings were obtained from independent ratings for three 35-s CD passages per hearing aid. The significant results of the Friedman's analyses (Siegel, 1956) of variance of these rankings indicated that the four hearing aids were ranked very similarly by most or all subjects. This provides strong presumptive evidence that essentially the same rankings would have been obtained on repeated testing of normal hearers. However, as mentioned earlier, several investigators have shown that the reliability of judgments of hearing-aid-processed speech is somewhat poorer for hearing-impaired persons than for normal hearers. It is possible, therefore, that more than three judgments per instrument would be required for a hearing-impaired individual.

Another aspect of the intelligibility rating task which requires investigative attention is the optimal structure of the CD passages themselves. The length of CD passages which have been used for judgments of hearing-aid-processed speech has varied from 10 (Gray and Speaks, 1978) to 65 s (Studebaker *et al.*, 1982). However, Studebaker *et al.* (1982) noted that subjects usually required 20–40 s to make their decision. Clearly, for clinical application, passages should be long enough to permit the subject to form an opinion, but not substantially longer. Perhaps three 20-s judgments produce as reliable a result as three 40-s judgments.

On a related matter, comments made by subjects during the progress of this investigation seemed to indicate that the 24 CD passages recorded by each talker were not inherently equivalent in intelligibility in spite of the *a priori* efforts made to equate them. Retrospectively, the long term rms spectrum was measured for each CD passage. The passages for a given talker were all interweaving across the frequency range. No basis could be found for a postulate that the passages varied in difficulty because of different spectral content. It would appear prudent, therefore, to empirically equate the intelligibility of CD passages to be used in hearing aid selection. This would have to be accomplished with normal hearers and, hence, would not necessarily equate the passages for each hearing-impaired individual. However, this procedure could be used to eliminate any passages with intelligibility clearly deviant from the mean value.

ACKNOWLEDGMENTS

This study was supported in part by Grant #NS15996 from the National Institutes of Neurological and Communicative Disorders and Stroke. Appreciation is expressed to Sheila Smith for the preparation of the stimulus materials.

Chial, M. R., and Daniel, S. W. (1977). "Hearing aid quality judgements by normal and dysacusic listeners," paper presented at Am. Speech Lang. Hear. Assoc. Conv., Chicago, IL.

Cox, R. M., and Studebaker, G. A. (1980). "Problems in the recording and reproduction of hearing-aid-processed signals," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. A. Studebaker and I. Hochberg (University Park P.,Baltimore, MD), Chap. 9, pp. 169–196.

Fry, E. (1968). "Fry's readability graph," J. Read. **11**, 513–516, 575–581.

Giolas, T. G. (1966). "Comparative intelligibility scores of sentence lists and continuous discourse," J. Aud. Res. **6**, 31–38.

Giolas, T. G., and Epstein, A. (1963). "Comparative intelligibility of word lists and continuous discourse," J. Speech Hear. Res. **6**, 349–358.

Gray, T. F., and Speaks, C. E. (1978). "Ability of hearing impaired listeners to understand connected discourse," J. Am. Aud. Soc. **3**, 159–166.

International Organization for Standardization. Recommendation R 532 (1966). "Method for calculating loudness level," Part II, 9–23, Switzerland.

Kirk, R. E. (1968). *Experimental Design: Procedures for the Behavioral Sciences* (Wadsworth, Belmont, CA).

Lawson, G. D., and Chial, M. R. (1982). "Magnitude estimation of degraded speech quality by normal- and impaired-hearing listeners," J. Acoust. Soc. Am. **72**, 1781–1787.

Nakatani, L. H., and Dukes, K. D. (1973). "A sensitive test of speech communication quality," J. Acoust. Soc. Am. **53**, 1083–1092.

Punch, J. L. (1978). "Quality judgements of hearing aid-processed speech and music by normal and otopathologic listeners," J. Am. Aud. Soc. **3**, 179–188.

Punch, J. L., and Parker, C. (1981). "Pairwise listener preferences in hearing aid evaluation," J. Speech Hear. Res. **24**, 366–374.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences* (McGraw–Hill, New York).

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

Speaks, C., Parker, B., Harris, C., and Kuhl, P. (1972). "Intelligibility of connected discourse," J. Speech Hear. Res. **15**, 590–602.

Studebaker, G. A., Bisset, J. D., VanOrt, D. M., and Hoffnung, S. (1982). "Paired comparison judgments of relative intelligibility in noise," J. Acoust. Soc. Am. 72, 80–92.

Winer, B. J. (1971). *Statistical Principles in Experimental Design* (McGraw–Hill, New York), 2nd Ed.

Zerlin, S. (1962). "A new approach to hearing aid selection," J. Speech Hear. Res. 5, 370–376.