Key Issues in Hearing Aid Selection and Evaluation

Denis Byrne*

Abstract
This article examines three key issues in hearing aid selection and evaluation. First, two studies are analyzed to determine the extent to which differences in frequency response characteristics result in differences in judgments of the intelligibility or pleasantness of amplified speech. Small to moderate frequency response differences usually did result in different judgments although occasionally large differences did not. When two dissimilar responses are equally intelligible, it is likely that the best response is something in between them. Second, evidence indicates that it is not always satisfactory to allow frequency response to be determined by prescribing gain at each frequency separately. This suggests, therefore, that separate frequency response prescription rules are needed. Third, signal audibility is discussed in the context of using simplified Articulation Index schemes to predict aided performance. It is demonstrated how wrong conclusions may easily be drawn in the absence of accurate knowledge of the signal levels received when different hearing aid options are used. Data suggest that the Articulation Index may not be applicable to steeply sloping high-frequency hearing losses. This needs further investigation before the Articulation Index can be recommended for hearing aid evaluations.

Key Words: Frequency response characteristics, Articulation Index (AI), signal audibility

This article is concerned with three of the many issues in hearing aid selection and evaluation. All are important and can be discussed in relation to recent data from the National Acoustic Laboratories. The first issue concerns the importance of differences in frequency response characteristics. Or, put differently, what are the smallest differences, as measured physically, that are likely to lead to significant differences in performance with a hearing aid or the acceptability of it? The second issue is the need for separate gain and frequency response prescription rules, and the third is the question of signal audibility when wearing a hearing aid, with particular reference to the application of Articulation Index type procedures.

IMPORTANCE OF DIFFERENCES IN FREQUENCY RESPONSE

Today, almost everyone accepts the general concept that people with different hearing impairments require different frequency response characteristics. However, there is considerable room for debate about how accurate frequency response selection needs to be. Is it sufficient to fit any response within a broad range or, alternatively, is a relatively precise fitting desirable? Our view of this issue is critical because important decisions hinge upon it. For example, how much trouble should we go to when trying to match a prescription? Or, is it worth bothering about real-ear measurements or using a computerized fitting system? Does it matter which selection procedure we use? Or, how fine do the adjustments need to be on the hearing aids we design or use? To examine this question, I have analyzed data from two National Acoustic Laboratories' (NAL) studies.

Study 1

In the first study (Byrne, 1986), I compared three or four frequency response characteristics for 11 subjects comprising 14 test ears. Figure 1 illustrates one set of characteristics. For some subjects, the responses differed somewhat less than those shown in the figure, but overall, they could be considered to be moder-
ately different frequency responses. (All 14 response sets are presented in Figure 1 [Byrne, 1986].) The responses were evaluated by paired comparison judgments of the intelligibility of speech in quiet, intelligibility in noise, pleasantness in quiet, and pleasantness in noise. For these four judgments, each response was compared with each of the other responses ten times. If one member of a pair was chosen eight or more times, then that was significant at approximately the 5 percent level. In the present context, the question is: How much do responses have to differ before they are judged to be significantly different with respect to intelligibility or pleasantness? This question was examined by identifying the response that was best for each subject and then determining whether or not each of the other responses was significantly different from the best response with respect to each of the four judgments. This type of analysis was first undertaken by Dillon (1985) who used this set of data. The metric he used to quantify the differences between responses is shown in Figure 2.

First, the responses to be compared were equated for gain averaged over the frequencies 500, 1000, 2000, and 4000 Hz. Then, the differences in gain at each frequency were added together, and the final figure calculated was the square root of the mean of the differences squared. The rms differences between the best response and each of the other responses were calculated, and then, for all judgments combined, distributions of rms differences were drawn up for the judgments that were significantly different and those that were not.

Figure 3 shows the distribution of rms differences when responses were judged to be significantly different and the distribution when responses were judged to be not different. Even the smallest differences that were studied, namely, rms differences of 3 or 4 dB, resulted in significantly different judgments more often than not. However, in a small proportion of cases, large rms differences (6–10 dB) did not result in significantly different judgments.

Later, I analyzed the same data set in more detail by using several other metrics. These are explained in Figure 4. The first metric is the Total Slope (TS). This is the difference between the gain at 400 Hz and the gain at 3000 Hz. This can be broken down into two components (i.e., a low slope and a high slope). The low slope is the difference between the gain at 400 Hz and at 1250 Hz, and the high slope is the difference between the gain at 1250 Hz and 3000 Hz. A fourth metric was introduced after observing that responses similar in total

---

**Figure 1** Example of the frequency response characteristics, for one subject, which were evaluated in Study 1, using paired comparison judgments. The codes (M5, M1, M3, N) refer to the procedures used to prescribe the responses (details in Byrne, 1986).

**Figure 2** Method used to calculate rms differences between two frequency responses. First, the two responses, 1 and 2a, are equated for four-frequency average gain by adjusting 2a to 2b. Then, the rms difference is calculated as shown by the equation.

**Figure 3** Method used to calculate rms differences between two frequency responses. First, the two responses, 1 and 2a, are equated for four-frequency average gain by adjusting 2a to 2b. Then, the rms difference is calculated as shown by the equation.
slope often differed considerably in the gain at frequencies between 400 and 3000 Hz. This metric, designated Mid-Range Signal (MRS), is calculated by comparing the gain of the response in question with the gain shown by a straight line connecting 400 Hz and 3000 Hz. Specifically, it is the gain differences at the three frequencies 1000, 1250, and 1600 Hz. The values used in the analysis were the differences between two responses, with respect to the several metrics, versus the difference between how often one response was preferred to the other for each of the four judgments.

Table 1 shows the correlations between the differences in responses and the differences in judgments. The first two rows show the correlations for total slope and rms. For each judgment, the correlations for these two metrics were similar. This is not surprising because, for these response sets, the total slope and rms differences were highly correlated with each other (i.e., a correlation of 0.89). Looking at the

![Figure 3](image1.png)

**Figure 3** Distributions of comparisons that were significant (one response preferred to the other eight or more times out of ten), and those that were not significant, as a function of rms differences between responses. For each subject, the best response was compared, using paired comparison judgments of intelligibility and pleasantness, with each of the two or three other responses that were tested.

![Figure 4](image2.png)

**Figure 4** Explanation of parameters used to describe frequency responses.

<table>
<thead>
<tr>
<th>Frequency Response Parameter</th>
<th>Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Slope (TS)</td>
<td>.45†</td>
</tr>
<tr>
<td>rms</td>
<td>.45†</td>
</tr>
<tr>
<td>Low Slope</td>
<td>.16</td>
</tr>
<tr>
<td>High Slope</td>
<td>.26</td>
</tr>
<tr>
<td>Mid-Range Signal (MRS)</td>
<td>.11</td>
</tr>
<tr>
<td>TS + MRS</td>
<td>.48†</td>
</tr>
</tbody>
</table>

*Judgments were for intelligibility in quiet (IQ), intelligibility in noise (IN), pleasantness in quiet (PQ), pleasantness in noise (PN).
†Significant at 1% level; ‡Significant at 5% level.
Table as a whole, the correlations are generally low and are only significant for total slope and rms and only for the intelligibility in quiet and pleasantness in quiet judgments. If we consider multiple correlations, as shown in the bottom line, the addition of mid-range signal to total slope increases the correlation from 0.45 to 0.48.

To examine the data more closely, scattergrams were constructed showing the relationships of response differences to judgment differences.

Figure 5 shows how differences in total slope were related to differences in intelligibility in quiet judgments. A judgment difference of six or more is significant as this means that one of the responses was chosen eight or more times out of ten. The main point to note is that, for the majority of comparisons, even the smallest frequency response differences produced significant judgment differences. There are some instances, however, where relatively large response differences have not led to differences in judgments. Possibly this may arise because the responses differed not only in slope but also in other respects, notably in the amount of mid-range signal. It is possible that the effects of slope differences tend to be offset by the effects of other differences and that the final result is that two responses are equally intelligible although physically different and probably different in quality. This supposition was supported to a limited degree by an analysis which showed that slope differences tended to have less effect if accompanied by large differences in mid-range signal than if there were differences in the slope parameter only.

Figure 6 Response variations that were tested in Study 2. The variations are expressed as the difference from a reference response (Response 1), which had been individually selected for each subject (for details of experiment see Byrne et al, 1990).

Study 2

The significance of response differences might be clarified by examining a study in which the responses varied in only one parameter. This has been done using the data of a study of the frequency response requirements of 46 severely and profoundly hearing-impaired subjects (Byrne et al, 1990). The subjects performed intelligibility judgments while wearing hearing aids and listening to samples of speech that had been filtered to simulate the effects of changing the hearing aid's response.

Figure 6 shows the variations from the base response that had been individually selected for each subject. If we represent the base response by a straight line, then Response 2 has 6 dB/octave greater slope from 250 Hz to 2000 Hz, whereas Response 3 has 6 dB/octave less slope and Response 4 has 12 dB/octave less slope. Each response was compared with every other response ten times. This means that, for each subject, there were three pairs of responses that differed by 6 dB/octave. These are numbers 1 and 2, 1 and 3, and 3 and 4. There were two pairs of responses which differed by 12 dB/octave. These are numbers 1 and 4, and 2 and 3. There was one pair of responses, numbers 2 and 4, which differed by 18 dB/octave. Thus, these data can be used to see how often slope differences of 6, 12, and 18 dB/octave resulted in significantly different judgments.
Figure 7 shows how often response pairs were judged to be different as a function of the slope differences. The subjects have been divided into two groups for which the results are shown separately. The reason is that, although intelligibility judgments were sought, 15 subjects reported that they could not understand anything with any of the responses. Those subjects were instructed to choose the response that they would prefer in a hearing aid. Considering the left panel of the figure, the results are as expected in that the proportion of significant differences in judgments increases as the difference in slope is increased. Note, however, that even the 6-dB difference usually resulted in a significant difference in judgments. For subjects in the right panel, the trends are less marked, indicating that differences in slope had less effect on differences in judgment. The probable reason for this is that the subjects in the quality group were not hearing the higher frequencies (Parkinson et al., 1989). As a result, there was actually less difference in the audible signal when these subjects listened with different frequency responses than there was for the intelligibility group.

Although the judgments were usually affected by a 6-dB difference, at least for the intelligibility group, there were some comparisons for which differences of 12 or even 18 dB/octave did not result in significantly different judgments. A possible explanation might be that the responses being compared were removed from the best response by similar amounts but in opposite directions. This possibility was examined using an estimate of the best response (i.e., the one that was preferred most times in total in the comparisons with the other three responses). If the above explanation is correct, then we would expect the best response to fall between the two responses that were judged to be equal. Specifically, if responses 2 and 4 (see Fig. 6) were equal, the best response should be either 1 or 3 and both of these responses should receive higher scores than either 2 or 4. That was, in fact, the case for all seven subjects who judged 2 and 4 to be equal. Similarly, if 2 and 3 were judged to be equal, it would be predicted that 1 was the best response, and if 1 and 4 were equal, it would be predicted that 3 was the best response. The same explanation could apply to some of the equal judgments of adjacent responses. For example, it may be that if 3 and 4 were judged equal, then the best response could be about halfway between the two. This explanation would be plausible if only one pair of adjacent responses were judged to be equal and if the scores for both responses were higher than either of the other two. Figure 7 indicates how many of the equal judgments could be explained by the hypothesis that the responses were similarly distant from the best response. As mentioned before, it accounts for all of the 18-dB differences. For the intelligibility group, it could also account for 10 of 11 of the 12-dB differences, and for the quality group, it could account for 9 of 13 of the 12-dB differences. It also could account for a significant proportion of the 6-dB differences.
Conclusions

It is clear, particularly from the first study, that small to moderate sized differences in frequency response often result in consistent differences in the perceived intelligibility or pleasantness of speech in quiet or in noise. This point is emphasized by the fact that, in the first study, there was only one comparison of the total of 47 that did not result in a significant difference for at least one of the four judgments, and there were only eight comparisons that did not result in differences for at least two judgments.

Provided that we regard such differences as important, it does seem that we should try to meet frequency response prescriptions relatively accurately. This supports the practices of doing real-ear measurements and of modifying aid fittings to obtain a more accurate match to a prescription. It also supports the need for aids that can be fitted accurately and particularly for improvements in areas that are deficient, notably in the gain above 3000 Hz.

Large differences in response sometimes do not result in differences in a particular type of judgment. They nearly always, however, result in some differences if a number of judgments are made of different qualities or using different materials. This is an argument for including more than one type of judgment when evaluating amplification systems. Two explanations have been suggested for why very different frequency responses may sometimes be judged to be equal. One is that the effects of differences in one parameter may be offset by differences in another parameter. However, even when responses differ in only one parameter, they may be judged to be equal because the best response is about mid-way between the two that were tested. This point is worthy emphasizing because we might be tempted to assume that anything between two equal responses would also be equal. In fact, the two responses may be equally bad, rather than equally good, and something in between may be much better.

Indeed, when two substantially dissimilar responses are judged to be equal, then we should suspect that neither of them is optimal.

SEPARATE FREQUENCY RESPONSE AND GAIN PRESCRIPTION RULES

The literature on gain and frequency response prescription rules is extensive (for reviews, see Byrne, 1983; Skinner, 1988; Byrne, in press). This article addresses the need for separate frequency response prescription rules in addition to gain rules.

In most prescriptive procedures, gain is selected at each frequency independently, and this determines the prescribed frequency response. Thus, if gain were prescribed to be half of hearing level at each frequency, then the difference in response slope between frequencies would be half of the audiogram slope. In some procedures, however, there are separate rules for prescribing the average gain and for prescribing frequency response. For example, the NAL procedure (Byrne and Dillon, 1986) uses a half-gain rule combined with a one-third slope rule. This is unusual, but there is at least one precedent, namely one of Lybarger’s procedures, which was a half-gain rule combined with a quarter-slope rule (Lybarger, 1953). There are also several procedures in which the frequency response is constrained in some way so that it is not always what would be obtained by applying the gain rule to all frequencies (examples cited later).

A basic issue concerns the need for separate rules for prescribing gain and frequency response. In other words, does the required gain at any particular frequency depend solely on the hearing loss parameters at that frequency or does it depend partly on the hearing loss or the required gain at other frequencies?

For indirect evidence in favor of a separate rule, we can use the Harvard report (Davis et al., 1947). That research indicated that large variations in audiograms required only small variations in frequency response slopes and may have been a factor in prompting Lybarger to change from his 1944 formula, which used a gain rule only (Lybarger, 1944), to his 1953 formula, which used separate gain and response slope rules (Lybarger, 1953). More direct and more extensive evidence is available from the research relating to the development and validation of the various prescriptive procedures.

Such evidence was obtained when evaluating the original NAL procedure (Byrne and Tonisson, 1976). One major finding was that the prescribed response slopes were too great, particularly for steeply sloping audiograms (Byrne, 1986). An analysis of the relationship of required response slope to audiogram slope, using data from two NAL and three other studies, indicated that response slopes needed to be varied by only a third as much as the variations in audiogram slopes (Byrne and Murray, 1986). The new NAL procedure (Byrne and Dillon, 1986), therefore, incorporated a one-third slope...
rule combined with a half-gain rule. The appropriateness of these rules was confirmed by a validation study (Byrne and Cotton, 1988). It would have been unsatisfactory simply to use a third-gain rule as that would have prescribed sufficient gain only for the milder hearing losses.

As mentioned earlier, there are several procedures in which the response slope is constrained by some rule in addition to the gain rule used in the procedure. The reasons for such constraints have not always been stated, but two instances where they have are the CID and MSU procedures described by Skinner (1988) and Cox (1988), respectively. Skinner recommends that the difference between the gain at 500 Hz and 2000 Hz should never exceed 35 dB. Cox recommends that amplified speech bands should always be at least 15 dB below the upper limit of comfortable loudness. Both of these constraints have the effect that, in some cases, typically those with steeply sloping hearing losses, the response slope will be less than would be obtained by applying the gain rule at all frequencies. The research, which supports these constraints, may also be regarded as evidence that separate slope rules are needed for at least some cases.

Recent research on the gain and frequency response requirements of the severely and profoundly hearing-impaired (Schwartz et al., 1988; Byrne et al., 1990) has led to modifications of both the NAL and the POGO procedures. Both procedures now change the gain rule when hearing level reaches 60 or 65 dB, and in this respect, they are similar. However, the two procedures differ substantially in the way that they handle frequency response. This is illustrated in Figure 8.

The left panel displays two audiograms of identical configuration but differing severity. The right panel shows the responses prescribed by NAL, as recently modified (Byrne et al., in press), and by POGO II (Schwartz et al., 1988). The point of this figure is to show how the POGO response changes with increasing severity of hearing loss and to compare this with how the NAL response changes with increasing severity of hearing loss. To facilitate the comparison of response shapes, the POGO response for the less severe loss has been raised by 20 dB, and the NAL response for the more severe loss has been lowered by 20 dB. If the gain rule is applied at all frequencies, as in POGO, the frequency response slope becomes steeper for the more severe hearing loss. By contrast, the NAL response is flatter for the more severe hearing loss. The reason is that the formula was derived from research that indicated that the required response slope becomes less steep when the hearing loss becomes very severe at the high frequencies (Byrne et al., 1990). Thus this research demonstrates that the simple application of a gain rule across all frequencies may lead to an inappropriate frequency response prescription. Specifically, the research indicated a need for separate slope and gain rules for the severely hearing-impaired and that both these rules needed to differ from what is appropriate for the less impaired.

Figure 8 Comparison of NAL (modified for severely/profoundly hearing impaired) and POGO II prescriptions for audiograms with same shape but differing severity. Note that, as severity increases, POGO increases the degree of high-frequency emphasis whereas NAL decreases the high-frequency emphasis. This illustrates one effect of using separate gain and response slope rules (NAL) compared with using a gain rule only (POGO).
Although much has already been written about gain prescription, it seems that we need to give more attention to frequency response rules because it is not safe to rely on gain rules to take care of frequency response selection automatically.

**SIGNAL AUDIBILITY**

Signal audibility has always been a key concept in understanding hearing aid fitting because it is generally regarded as the most important, although not the only, determinant of speech intelligibility. Signal audibility considerations have become especially prominent with the strong current interest in using the Articulation Index (AI) or similar schemes to evaluate the options for hearing aid fitting. Such theoretical models have enormous potential (Studebaker, 1991), and at least three simplified AI schemes have been proposed for clinical use (Popelka and Mason, 1987; Pavlovic, 1988, 1989; Mueller and Killion, 1990). Signal audibility will be discussed within the framework of simplified AI schemes but with the intention of making a number of more general points.

Thanks to the efforts of the above authors, among others, it is easy to perform AI calculations for unaided or aided listening conditions. However, it is also easy to draw the wrong conclusions if the AI is applied without a good understanding of the factors that affect signal audibility under realistic conditions of hearing aid use. For example, there is no point in calculating what a person ought to be able to hear with a volume setting that he or she will not accept. Similarly, there is no sense in comparing different hearing aid options unless we have accurate information on how much audible signal each would provide if they were used. These basic facts have been ignored in some hearing aid studies using the AI, with the result that the conclusions drawn are meaningless or questionable. In fact, most calculations of signal audibility involve some assumptions, and the real issue is, therefore, to ensure that these assumptions are sufficiently reasonable to permit useful conclusions. The following sections consider factors that affect signal audibility and evaluate some of the assumptions that have been made, or could be made, when calculating AIs.

**Factors Affecting Signal Audibility and AI Calculations**

In Figure 9 signal audibility is considered in the most simple listening condition (unaided).

These graphs are in the format of the simplified procedure of Pavlovic (1988). The essential elements are the thresholds of the subject and two lines representing the peaks and minima of speech. The AI is calculated by adding how much the speech peaks exceed thresholds at each frequency, with the proviso that the maximum contribution is limited to 30 dB. Because we have measured the thresholds, there is no doubt where they should be placed on the graph. We could, however, argue about the position of

![Figure 9 Illustration of simplified method for calculating Articulation Index (AI), as proposed by Pavlovic (1988), and (right panel) effect of using a higher speech level.](image)
the speech peak line which, as shown in the left panel of the figure, assumes an overall speech level of 63 dB SPL. If we assumed a lower speech level or a higher level, then the line would move up or down the graph correspondingly and this would change the AI. Other authors have suggested different speech levels. For example, in one scheme used in NAL, that level was 70 dB SPL (Byrne, 1978). The right panel of the figure shows the AI that results from using a level of 70 dB instead of 63 dB. The AI, which is proportional to the shaded area on each graph, increases from 0.17 to 0.38. The choice of speech level is rather critical because even a difference of a few dB usually adds or subtracts a few dB across all or most frequencies and, thereby, has a substantial effect on the AI. The choice of speech spectrum, represented here by a straight line, is also an issue, but the effect of using different estimates would be small.

What then is the most appropriate speech level? In most AI schemes, the rationale has been to choose a level that is representative of normal conversation. Although this is logical, I would argue that, more commonly, communication involves interaction of the aid wearer with other people, and when this happens, the other people tend to adjust their voice levels to the extent needed for good communication. This is usually above a normal level and, presumably, is required to provide an improvement in signal-to-noise ratio as well as a higher level.

AI calculations, for aided conditions, should be based on a level that is typical of what the hearing aid wearer would receive rather than on a normal conversational level. There are several studies that permit inferences about the typical speech input levels received by hearing aid wearers. These studies suggest that, for mild to moderately impaired listeners, a typical level would be between 65 and 70 dB SPL (Martin et al., 1976; Walden et al., 1977; Farrell et al., 1979). Figure 10 shows data of this type from two NAL studies. The data points, except for the single enlarged point, are from a study with (mostly) mildly to moderately impaired subjects (Byrne and Cotton, 1987). The regression line shows that, as hearing level increases, the typical speech input level also increases. Although there is a lot of scatter of the data points, the trend is significant at the 5 percent level. It indicates that for a 100-dB increase in hearing level, there is a 7.5-dB increase in speech level. The enlarged point on the far right of the figure is the average of data from a study with severely and profoundly impaired subjects (Parkinson et al., 1989). This was not used in the regression analysis but would seem to confirm the trend that is shown. These data suggest that, if we use the AI over a large range of hearing losses, we should consider using two or more speech input levels depending on the degree of hearing loss. The appropriate level would appear to be between 65 and 70 dB SPL for a moderate hearing loss and between 70 and 75 dB SPL for a severe loss.

Figure 11 shows how the AI is calculated for aided listening.
considerable room for error because rather small differences in overall gain can make large differences in the AI. In the example shown in Figure 11, if we assumed 3 dB too much gain, the AI would be changed from 0.73 to 0.83, whereas it would change to 0.63 if we assumed 3 dB too little gain. One assumption that certainly is not safe is that the client will turn the volume up to maximize the AI. In fact, NAL data on preferred listening levels suggest that the AI will rarely, if ever, be maximized for clients with more than a moderate hearing loss, and even for mild losses, it will not always be maximized (Byrne and Cotton, 1987).

Theoretically, loudness discomfort level is another complicating factor in AI calculations. However, NAL data on preferred listening levels suggest that discomfort level will rarely enter the picture if the calculations are based on levels that hearing aid wearers will use. This is not surprising considering that the individual speech band levels would need to be well below loudness discomfort level to avoid discomfort for broad-band sounds.

There is, however, another factor that is important for severely impaired clients and possibly for others. The amount of signal received when aided may be limited by the aid's saturation sound pressure level (SSPL) to something less than the value obtained by adding the aid's gain to the unaided speech level. In a study of severely and profoundly hearing-impaired listeners, we calculated the sensation levels of the speech peaks received by the subjects with

![Figure 11](image1.png)

**Figure 11** Calculation of AI for aided listening. Note change in AI that would result from assuming 3 dB too much or 3 dB too little gain.

The basic difference is that the speech level is raised by the gain of the hearing aid. (In this example, various amounts of gain, at different frequencies, have been added to the unaided speech peak levels shown in Figure 9.) This procedure is fine with one vital proviso (and assuming resolution of the previously discussed issues, notably the choice of speech level). The proviso is that we need to have measurements of, or an accurate means of predicting, the gain that will be used. Unfortunately, several studies have proceeded without such knowledge and with the consequence that the conclusions drawn are a direct reflection of the assumptions that were made. For example, some studies (Berger, 1990; Rankovic, 1991) have compared procedures in terms of the gain prescribed and have, predictably, come to the conclusion that the procedure that prescribes the most gain will produce the highest AI. This does not, however, give any indication of the relative amounts of signal that would be received as a consequence of using different prescriptions. It fails to consider that when a hearing aid is used the hearing aid wearer will adjust the overall gain to the preferred level regardless of what is prescribed. (Similarly, performance test measurements may be misleading if comparisons are made at levels derived by adding differing amounts of prescribed gain to a fixed presentation level.) If we have to make assumptions about gain, there is
their aids on the best tone setting and preferred volume setting (Parkinson et al., 1989). We found that for 12 of 46 subjects, the sensation level was limited by SSPL for at least one, and usually more of the frequencies 500, 1000, 2000, and 4000 Hz. An example of this is shown in Figure 12.

This figure shows the subject's thresholds, the real-ear SSPL, and the speech peak levels that would have been achieved if they were not limited by SSPL. The AI has been calculated, by Pavlovic's method, both including and excluding the SSPL limitations. We see that the speech levels are limited by SSPL at 500 Hz and 1000 Hz, and this has reduced the AI from 0.48 to 0.42. There is also a slight reduction at 4000 Hz but this has had no effect because the speech peaks were below threshold. In addition to the AI issue, this example illustrates another important reason for including SSPL on graphs of this type. If we were not aware of the SSPL limitation, we might suppose that the provision of more gain could be beneficial, particularly at 4000 Hz. However, it is clear that we could not get any extra signal by providing more gain unless SSPL were also increased.

**Frequency Response Comparisons Using AI**

The complications of applying the AI are reduced when it is used in a relative manner, such as estimating which of two frequency response characteristics would be better, rather than when it is used to predict absolute levels of speech recognition. Nonetheless, there are many pitfalls that can lead to incorrect conclusions. This is illustrated with a series of examples (Figures 13–16) from a study in which I determined the preferred listening levels for different frequency response characteristics. These examples compare the AIs calculated from the measured speech levels with AIs based on various assumptions.

Figure 13 shows the speech peak levels for two frequency responses after the speech had been adjusted to the preferred level for each response. The AIs have been calculated by the method described by Pavlovic (1989), using eight frequencies rather than the simplified method. Response 1 provides more signal in the high and low frequencies, compared with Response 2, which provides more signal at the mid-frequencies. This illustrates the general point that when we change the frequency response and allow the client to readjust the gain to the preferred level, we virtually always finish up with one response providing more signal at some frequencies and the other response providing more signal at other frequencies. This is because the client will select a constant loudness level or something similar to that. The evaluation question therefore concerns whether the additional signal at certain frequencies is worth more or less than the accompanying reduction in signal at other frequencies. This issue requires consideration of the importance attached to different frequencies. In this example, if we regarded the midfrequencies as particularly important, we might conclude that Response 2 was better. We might conclude that it was poorer if we regarded the midfrequencies as unimportant. The frequency importance function is an issue because at least two simplified procedures (Popelka and Mason, 1987; Mueller and Killion, 1990), as well as most research applications (e.g., Humes, 1986; Dugal et al., 1980), have used the values for predicting understanding of syllables or single words whereas Pavlovic has proposed, very logically, that we use a set of frequency importance values that are appropriate for average speech. This is an issue that needs to be resolved. However, it will not be pursued here because, for the data from which these examples were drawn, it did not make much difference whether the AIs were calculated with the Pavlovic or the nonsense syllable importance functions. Logically, how-
However, the choice of importance function could become significant when comparing some types of frequency responses, notably if one provided considerably more low-frequency signal and considerably less high-frequency signal.

In Figure 13, note that the AIs are the same for both responses, namely 0.36. This example is based on measurements of the subject's preferred listening levels. However, the following figures will show what happens if we do not have such measurements and we need to make assumptions about signal levels. One assumption that has been made in some examples in the literature is to equate the gain of the two responses at one frequency, such as 2000 Hz (Fig. 3 of Mueller and Killion, 1990). We could do this, for this example, by lowering the gain of Response 2 by 3 dB.

Figure 14 shows Response 1 at the preferred level but Response 2 reduced by 3 dB to equate the gain at 2000 Hz. Although this gain adjustment is not large, it has substantially reduced the AI for Response 2, namely from 0.36 to 0.26. If we compared the gain-eqeated responses, we would conclude that Response 1 was clearly superior, whereas, in fact, the AIs are equal if the comparison is made at the preferred listening levels. This example shows that it is critical to be accurate in estimating the relative levels of the responses being compared. This needs to be known within 1 or 2 dB; otherwise, there is a strong likelihood of drawing the wrong conclusions.

The next example shows why it can be important to know the absolute signal levels rather than just the difference in levels provided by different responses. To illustrate this, the AIs have been calculated for the two responses with each of them lowered by the same amount, namely 6 dB. This is shown in Figure 15. The AIs are considerably reduced for both responses, but the point to note is that the AI for Response 2 has been more reduced than the AI for Response 1. If we compared the responses at these levels, instead of at the preferred levels, we would once again erroneously conclude that 1 was superior to 2. The reason why one AI is reduced more than the other is that, in the mid-frequency range, the signal is reduced to zero, but this happens after only a 2- or 3-dB gain reduction for Response 1 whereas it takes a 6-dB reduction for Response 2. Consequently, the AI contributions for the midfrequency bands are reduced by the whole 6-dB gain reduction for Response 2 whereas only the first 3-dB reduction has any effect on Response 1. We can
Figure 16 Example of type of listener for whom AI may not be applicable. Although Response 1 has the highest AI out of four responses which were compared (two not shown), it was consistently judged to be the least intelligible.

also get misleading AI results if we assume an incorrectly high overall signal level. This may happen if the AI contributions are limited by the signal reaching the maximum of 30 dB or by reaching discomfort level, or a hearing aid's saturation level.

AI Applicability

It was not intended that this article address the validity of the AI as such. However, when working through examples, it was striking how often the AI predictions disagreed with speech intelligibility judgments, which had been obtained from these subjects.

Figure 16 presents an example of such a case. Response 1 will give the higher AI because it provides more signal at the high frequencies. (Although there are large signal differences in the low frequencies, these do not count because both responses provide more than 30 dB.) The point of interest is that, although the highest AI was obtained with the response with the most high frequency emphasis, that response was consistently judged to be the poorest out of a total of four responses that were compared.

This was true for judgments of intelligibility in quiet and also for intelligibility in noise and pleasantness in quiet and in noise. In a sample of 14 ears (11 subjects), there were seven with ski slope hearing losses similar to this one. In all these cases the response with the most high frequencies produced the highest AI. However, in six out of seven ears, that response was judged to be the poorest of a set of either three of four responses (Byrne, 1986). This example is presented as a caution that there may be some types of cases where the AI predictions break down even though the AI has been calculated with an accurate knowledge of signal levels.

This supposition is supported by the study of Rankovic (1991) who found that the subjects who had sloping high-frequency hearing losses did not achieve the best speech recognition scores with a frequency response that maximized the AI.

Conclusions

First, the levels at which hearing aid wearers choose to listen, would place the speech peaks about the middle of the dynamic range, rather than close to LDL. For hearing aid selection purposes, we may need to revise our concept of dynamic range because it appears that, as currently conceived, the upper half, or at least upper third, of the dynamic range represents signal levels that are not acceptable for long-term listening.

Second, preferred listening levels are usually not sufficient to maximize the AI for clients with more than mild hearing losses.

Third, in order to permit any valid conclusions, AI calculations need to be based on accurate estimates or measurements of the signal levels received in real listening situations. If the purpose is to compare different frequency responses, then their relative levels need to be known very accurately. Furthermore, absolute levels may also be important. Therefore, we need accurate information about gain levels, and we need to standardize and validate the overall speech level, which may need to be varied for different degrees of hearing loss. Although less influential, the speech spectrum and frequency importance functions should also be standardized.

Although the general validity of the AI is well established, there may be some types of hearing-impaired clients for whom it is not applicable. My data suggest that this may be true for steep, high-frequency hearing losses. This needs to be investigated further before recommending the AI for general use in hearing aid evaluations.
Finally, I make the observation that, when frequency responses are compared at the preferred listening levels, the differences in the AI are usually small. If we are to conclude anything from such differences, we must ensure that they represent the differences in signals that would be obtained under actual conditions of hearing aid use.

Acknowledgment. My thanks to Harvey Dillon for including his analysis of the relationship of rms differences in frequency responses to differences in judgments.

REFERENCES


