ISSN 0280-6819

TECHNICAL AUDIOLOGY

Report TA No. 112
November 1985

THE EFFECTS OF DIFFERENT FREQUENCY RESPONSES ON SOUND QUALITY JUDGMENTS AND SPEECH INTELLIGIBILITY

Alf Gabrielsson, Bo N. Schenkman and Björn Hagerman

ISSN 0280-6819
Report TA112
November 1985

Karolinska Institutet
Dept. of Technical Audiology
KTH
S-100 44 Stockholm
Sweden
Tel: +46-8-11 66 60

THE EFFECTS OF DIFFERENT FREQUENCY RESPONSES ON SOUND QUALITY JUDGMENTS AND SPEECH INTELLIGIBILITY

Alf Gabrielsson, Bo N. Schenkman and Björn Hagerman

ABSTRACT

Four speech programs and two music programs were reproduced by five systems with different frequency responses. One of the responses was flat, the others represented combinations of reductions at lower frequencies and/or increases at higher frequencies. Twelve hearing impaired (HI) and eight normal hearing (NH) subjects listened monaurally to the reproductions at comfortable listening level. They judged the perceived sound quality on seven perceptual scales and a scale for total impression. They also rated how an ideal reproduction would sound. Speech intelligibility under the different reproductions was measured by PB words and by sentences in noise.
The sound quality ratings showed good or satisfactory reliability. The multiple correlation of all perceptual scales with the total impression was high. Significant differences among the reproduction systems appeared in all scales for the NH listeners and in most scales for the HI listeners. The best system was characterized by a flat response below 1 kHz and a 6 dB/octave increase from 1 to 4 kHz. Rating ideal values worked well for the NH listeners but worse for the HI subjects. The PB words did not differentiate among the systems, and the S/N threshold for the sentences in noise only distinguished the flat response as worse than all others. On the whole there was little correspondence between the results of intelligibility measures and sound quality measures. The latter provided more information and distinctions among systems.

This work was supported by the Swedish Council for Planning and Cooperation of Research, The Bank of Sweden Tercentenary Foundation, and the Swedish Ministry of Health and Social Affairs, Delegation for Social Research.

Table of Contents

1 INTRODUCTION

Methods for selection and fitting of hearing aids may be divided into two categories: methods based on measures of speech intelligibility and methods relying on perceived sound quality. The former category is older and still the most common. The latter category emerged, when it became obvious that measurement of speech intelligibility was not sufficient. People with hearing aids selected according to the intelligibility criterion often complained about unpleasant and unnatural sound quality. The sounds were described as "hard", "sharp", "metallic", "noisy", "strange" etc, and as a consequence many people refrained from using their hearing aid.

This conflict between intelligibility and sound quality is discussed by several authors. The Harvard study and the Medical Research Council study of 1947 (see Levitt, 1978, for a review) suggested as one alternative for amplification that the frequency response should rise, e.g. by 4 to 6 dB/octave between 300 and 4000 Hz. Although this may improve speech intelligibility, it also tends to make the sounds "sharp", "shrill", "screaming" and "irritating" (Gabrielsson & Sjögren 1979a). Thompson & Lassman (1969, 1970) found that listeners with high-frequency sensorineural hearing loss preferred a flat frequency response, although they reached higher intelligibility scores with a rising frequency response. Harris & Goldstein (1979) found practically no correlation between quality judgments and speech discrimination in a reverberant room or in a sound suite, neither for hearing impaired nor for normal hearing listeners. More recently Harris & Goldstein (1985) showed that quality judgments by means of magnitude estimations were far more effective and reliable for differentiating among hearing aids with similar electroacoustic characteristics than speech discrimination data.

Punch and his co-workers (Punch, 1978; Punch & Beck, 1980; Punch et al., 1980; Punch & Parker, 1981) emphasized the importance of the low frequency region for quality judgments of hearing aid processed speech. Extended low-frequency response increased listeners' preferences and correlated highest with the dominant perceptual dimension in a multi-dimensional scaling analysis. It was suggested that disagreement between quality judgments and intelligibility scores may depend on upward spread of masking from low frequencies. Although lowered cut-off frequency was preferred, it may degrade speech intelligibility by such masking. Similar arguments were made by Harford & Cox (1978), who found that intelligibility scores increased, if frequencies below 1500 - 2000 Hz were attenuated and the high frequency region was extended to 6500 Hz. This condition was also slightly preferred in noisy situations but not in quiet situations. With regard to music listening, Franks (1982) found that both normal hearing and hearing impaired listeners preferred an extended low-frequency region. Normal hearing listeners also preferred an extended high frequency region, while no such preference appeared among the hearing impaired listeners. Franks suggested that the hearing aid should have a convenient switching mechanism to allow the user to include low frequencies when listening to music, and to exclude them (thereby avoiding masking effects) when listening to speech.

Judgments about sound quality in the above-mentioned reports were usually made by simply stating which reproduction was preferred to another. However, perceived sound quality is a multidimensional phenomenon, that is, it is composed by a number of separate perceptual dimensions. A mapping of these dimensions would increase our knowledge and understanding of sound quality and its relations to other variables. For this purpose an extensive series of experiments was performed (Gabrielsson 1979a, Gabrielsson & Sjögren 1979a, 1979b), in which listeners judged the sound reproduction of loudspeakers, headphones, and hearing aids for different programs including music, speech, and sounds from everyday life. The judgments were similarity ratings and adjective ratings, analyzed by multidimensional scaling and factor analysis, respectively. Free verbal descriptions were also used. Eight perceptual dimensions were identified: clarity, fullness, brightness vs. dullness, hardness/sharpness vs. softness, spaciousness, nearness, extraneous sounds, and loudness. Overall evaluations, such as "fidelity" or "pleasantness" may be considered as weighted functions of these perceptual dimensions. The physical correlates of the dimensions were investigated by studying the frequency response and the distortion of the systems and in some cases by manipulation of these variables. The principles underlying this approach have also been discussed by Gabrielsson (1981).

The perceptual dimensions were used in a clinical study on sound quality of hearing aids (Gabrielsson et al., 1980). Twelve hearing impaired subjects rated the sound quality of three aids, which were preliminary chosen to fit each subject's impairment (different aids for different subjects). Nine rating scales were used, one for each of the eight perceptual dimensions, plus a scale for the "total impression" of the sound quality. The scales were graded in five steps. The stimuli included three music programs, three speech programs, and four programs with everyday life sounds (e.g., traffic noise, typewriter). They were presented over a loudspeaker, and the respective hearing aid was adjusted to a comfortable listening level. The subjects visited the clinic at least four times at an interval of 2-3 weeks between them. At each occasion they rated one of the aids, then used it until the next visit, when another aid was rated etc. Furthermore repeated ratings were made of the earlier used aid(s), and supplementary ratings were done during the weeks that the aid was used in real life.

The results indicated satisfactory reliability of the ratings and confirmed the validity of the scales. Half of the subjects chose the aid, which they rated highest in "total impression", and the remaining subjects chose the aid rated next highest (and this rating was not significantly different from that of the highest rated). The whole procedure was very time-consuming, and it was suggested that the sensitivity of the scales could be increased by using more grades and more detailed instructions. Jerlvall et al. (1983) used the same scales for perceived sound quality as part of a larger study comparing in-the-ear hearing aids with behind-the-ear aids. Two thirds of the subjects chose the hearing aid, which they rated highest with regard to sound quality. However, speech discrimination scores were better correlated with the choice of aid than were the sound quality ratings.

Obviously both speech intelligibility and perceived sound quality are important criteria for the selection and fitting of an appropriate hearing aid. However, their relative importance and their relationship to each other depend on many factors, such as the specific way of measuring intelligibility and sound quality, which dimensions of perceived sound quality are considered most important, the type of sounds used as stimuli (speech, music, noise etc.), the type and degree of hearing loss, and the individual's listening habits. Some of these questions are addressed in this paper. More specifically we wanted to

(a) continue the studies on ratings of perceived sound quality by hearing impaired listeners,

(b) further investigate the effects of different frequency responses on perceived sound quality and on speech intelligibility, and

(c) thereby also contribute to clarify the relationsship between intelligibility and perceived sound quality.

The present investigation was therefore conducted, in which the frequency response was systematically varied for reproductions of speech and music. Speech intelligibility was measured by phonetically balanced words and sentences heard in noise (Hagerman, 1984a). The rating scales for sound quality were adapted from recent experiments on loudspeakers reproductions (Gabrielsson, Frykholm & Lindström, 1979; Gabrielsson & Lindström, 1981; Gabrielsson, Lindström & Elger, 1983, summarized in Gabrielsson & Lindström, 1985). The data for the hearing impaired listeners were treated individually as well as for the whole group. A group of normal hearing listeners was also included to allow comparisons.

2 METHODS

In summary the methods meant that 12 hearing impaired and 8 normal hearing subjects listened to 4 speech programs and 2 music programs, reproduced by 5 different frequency responses, and rated the sound quality on 7 perceptual scales and a scale for overall evaluation. Two of the speech programs were also used to measure speech recognition.

2.1 Stimuli, listening conditions

2.1.1 Programs

Intelligibility was measured by equivalent lists of phonetically balanced (PB) words and equivalent lists of sentences heard in noise (Hagerman, 1982, 1984a, 1984b).

There were six programs for quality judgments:

1. Female voice reading a fairy-tale.

2. Male voice in a disturbing background of other voices.

3. A list of 50 PB words originally used for measurement of speech recognition. (Silent intervals were deleted.)

4. A list of sentences heard in noise originally used for measurement of speech recognition in noise. (Silent intervals were deleted.) The signal-to-noise ratio was fixed to +10 dB.

5. Jazz music, excerpt from "Ole Miss" by W.C. Handy, performed by the Peoria Jazz Band in an auditorium. Gramophone record: OPUS 3, 79-00, Testskiva 1: Perspektiv.

6. Female solo voice, the folk chorale "Fröjda Dig, Du Kristi brud", performed by Marianne Mellnäs in Oscar Church in Stockholm. Gramophone record: Proprius PROP 7762.

These six programs lasted for about one minute each. Programs 1-2 were earlier used in Gabrielsson et al. (1980) and programs 5-6 in Gabrielsson, Lindström & Elger (1983).

2.1.2 Reproduction systems

All programs were monophonically recorded and played back on a tape recorder, Revox B77, 19 cm/sec. For the measurements of the speech reception threshold in noise the speech and the noise were recorded on separate channels and mixed to the desired signal-to-noise ratio during the measurement. The signal (program) from the tape recorder was fed into one out of five electrical filters. All filter settings were low-passed at 8 kHz. The electrical responses of the five filters were:

1. Flat

2. Flat below 1 kHz and increased 6 dB/octave 1-4 kHz; here abbreviated (flat, +6).

3. Attenuated 6 dB/octave below 1 kHz and increased 6 dB/octave 1-4 kHz, (-6, +6).

4. Attenuated 12 dB/octave below 1 kHz and increased 6 dB 1-4 kHz, (-12, +6).

5. Attenuated 12 dB/octave below 1 kHz and flat above 1 kHz, (-12, flat).

The schematic electrical responses of the filters are illustrated in Figure 1.

Insert Figure 1 about here

These frequency responses were chosen (1) because of previous recommendations on response curves in hearing aids (cf. Levitt, 1978), and (2) in order to study the effects of different responses below and above 1 kHz independent of each other.

There were thus three different responses below 1 kHz (flat, attenuated 6 dB/octave, attenuated 12 dB/octave) and two responses above 1 kHz (flat, increased 6 dB/octave). This gives six possible combinations, five of which were included here. The excluded one is that which would be attenuated 6 dB below 1 kHz and flat above. Attenuated treble is not included, nor is boosted bass, since these conditions are rarely used in hearing aids.

When measuring the speech reception threshold in noise, the noise channel was attenuated by one attenuator, and the program channel by another, both Hewlett & Packard, type 250D. The signals of the two channels were mixed and routed through one of the five filters. For all other stimuli conditions the noise channel was attenuated 100 dB.

The output was fed to a Western Electric 711A headphone. The response curve of the total system (filter + headphone), as measured with a Bruel and Kjaer coupler, type 4153, is illustrated in Figure 2. The so-called flat frequency response is sloping below about 200 Hz due to the features of the headphone. This headphone was chosen rather than the more common TDH 39 in order to avoid the resonance peak between 3 and 4 kHz.

Insert Figure 2 about here

The experiments were conducted in sound isolated chambers used for audiological or experimental purposes.

2.2 Subjects

There were 12 hearing impaired subjects, 6 males (age: 56, 48, 40, 29, 24, and 22 years) and 6 females (age: 57, 56, 49, 26, 24, and 19 years). They were attending a rehabilitation program of several months at a rehabilitation center. Their type of impairment is shown in Table 1. All of them worked in different occupations and had got hearing aids. Their opinions about the aid differed widely from "very good" to "useless" or "I never use it".

In the experiment the subjects listened monaurally with their best ear. The corresponding audiograms appear in Table 1. The hearing loss was usually mild or moderate at lower frequencies and increased more or less steeply toward higher frequencies.

There were also 8 normal hearing subjects (screening test 20 dB 125 - 8000 Hz), 4 males (18 - 30 years old), and 4 females (18 - 20 years old), all students.

All subjects were paid for their participation.

Insert Table 1 about here

2.3 Response variables

Sound quality was rated in 8 scales. Seven of these refer to perceptual dimensions: fullness (Swedish: fyllighet), loudness (ljudstyrka), brightness (ljushet), softness (mjukhet), nearness (närhet), spaciousness (rymdkänsla), and clarity (tydlighet). The English translations here differ somewhat from those used in earlier reports (Gabrielsson & Sjögren 1979a, 1979b). Translation into another language always causes some problems, and it should be remembered that the results refer to the scales defined by the Swedish labels. (Softness is used as an opposite to sharpness; an alternative translation might be gentleness.) The eighth scale was an overall evaluation, "total impression" (Sw. totalintryck).

Each scale was graded from 10 (maximum) to 0 (minimum), and with special definitions attached to 9, 7, 5, 3, and 1 as seen in Fig. 3. Decimals were included, since many subjects in earlier investigations used decimals in their ratings. Beside the definitions of the scales on the response form further explanation was given in the instructions, see Appendix 1. The use of these scales draws heavily upon the experience from sound quality ratings of high fidelity loudspeakers (Gabrielsson & Lindström, 1985).

Insert Figure 3 about here

Intelligibility was measured by equivalent lists of phonetically balanced words (PB) and equivalent lists of sentences heard in noise (Hagerman, 1984a). For each filter setting the subject listened to a list of 50 PB monosyllabic, Swedish words, preceeded by the carrier phrase "Now you hear...". The number of correct answers was noted.

The speech reception threshold in noise was also measured for each filter setting. Two lists of sentences were presented. The noise level was chosen to give somewhat more than 50% correctly recognized words in one list (out of the 50 words in the list, 10 sentences x 5 words in each), and somewhat less than 50% in the other list. Then the signal-to-noise (S/N) ratio corresponding to 50% correct recognition was calculated by linear interpolation between the percentages obtained from the two lists.

2.4 Design and procedure

Each subject took part in four listening sessions. The first session was devoted to determining comfortable listening level, the second to ratings of sound quality, the third to measures of intelligibility, and the fourth and last session to repeated ratings of sound quality plus answering a questionnaire. Each session lasted 1.5 - 2 hours, varying with the task and the individual, and one or two breaks were made. The interval between sessions varied from one day to about a week, sometimes more.

In the first session a comfortable listening level (Swedish "lagomnivå") was determined for each of the 30 stimuli (6 programs x 5 frequency responses). For each stimulus the experimenter started with a certain level, the subject listened for about 10 seconds and gave his opinion. Depending on this the experimenter either increased or decreased the level, and then continued the presentation. The subject listened and judged again, the experimenter made another adjustment, and so on, until a range for comfortable loudness was established. This usually required 2-3 minutes listening to each case. (Since each stimulus lasted about 1 minute, it had to be repeated once or twice.) The steps in the experimenter's adjustments were first large (up to 10 dB) and then decreased successively depending on the subject's judgments. The mean value of the range limits was used as listening level for the respective stimulus in the following sessions. The order of the filtered reproductions within each program was randomized as well as the order of the programs, differently for each subject.

In the second session the subject judged the perceived sound quality for each of the 30 stimuli according to the instructions in Appendix 1 and using the response form shown in Fig. 3. The presentation order of the filtered reproductions within each program was randomized as well as the order of the programs, differently for each subject. The order of the scales on the response form was different for different subjects but always the same within each subject. In connection with the instruction 12 preliminary trials were made as practice.

After all 30 stimuli were rated, the subject was asked to make ratings of ideal reproductions. As seen in the instructions for this task (Appendix 1), the subject again listened to the respective program but the task was to rate how the program should sound in each of the perceptual scales (fullness, loudness etc.) in order to give an ideal reproduction. It was then possible to compare the ratings for the real and for the ideal (imagined) reproductions. This procedure was used in several experiments on loudspeaker reproductions (Gabrielsson & Lindström, 1985).

In the third session the intelligibility tests were made in the way described earlier at the comfortable speech level. Five PB word lists and 10 lists of sentences in noise were used. The list used for each filter was randomly chosen, as well as the order of the filters. (The PB words were not used for the normal hearing subjects, since all of them would probably have obtained 100% correct answers.)

In the last session the subject again made ratings of the perceived sound quality and of the ideal reproductions. These data were then used to check the reliability of the subject's ratings. As usual the presentation order of the stimuli was differently randomized for each subject. Six preliminary trials were made. Finally the subject answered a questionnaire with various questions about the experiment. He/she was also asked to rank order the rating scales with regard to their importance for fidelity of the reproduction, and to give free and spontaneous comments on anything in the whole experiment.

2.5 Data treatment

The statistical treatment of the data generally followed the principles described in Gabrielsson (1979b, 1981) recently applied to judgments of loudspeaker reproductions (Gabrielsson & Lindström, 1985). Analysis of variance was used to estimate the reliability of the subjects' responses and to study the effects of different factors. Correlation and regression analysis provided a means for investigating the relationships among several variables; in some instances such relationships were also explored by means of factor analysis (component analysis).

3 RESULTS

3.1 Comfortable listening level

In average over subjects and programs the comfortable listening level was 16-20 dB higher for the hearing impaired group (HI) than for the normal hearing group (NH). Although there was a considerable interindividual variation in many respects, the level settings for the five reproduction systems showed a similar rank order for all subjects, HI as well as NH. For the HI group the order from highest to lowest level setting was system No. 5 (-12, flat), No. 4 (-12, +6), No. 3 (-6, +6), No. 1 (flat), and lowest No. 2 (flat, +6). In the NH group the order was the same with the exception that systems Nos. 3 and 4 were tied in rank. The difference between the extremes, Nos. 5 and 2, was usually 7-11 dB for individual HI subjects as well as for individual NH subjects. The rank order clearly indicates that reduction of the energy below 1 kHz meant a decrease in perceived loudness, and conversely that increased energy above 1 kHz resulted in higher loudness - as could be expected.

If we modify the schematic electrical response of the filters in Figure 1 to also reflect the mean differences in level settings among the systems, the result may be described as in Figure 4. The level setting for system No. 1 (flat) is used as reference (0 dB), and the schematic frequency responses of the remaining systems are adjusted vertically to reflect the difference in sound level in relation to the reference system. For instance, system No. 2 was set 3 dB lower than the reference in the HI group, while system No. 5 was set about 7 dB higher. As seen in Figure 4 the modified response curves are very similar for both listener groups.

Insert Figure 4 about here

A correlation analysis was performed in order to study the HI subjects' comfortable listening level as a function of the hearing loss at different frequencies. For the HI group as a whole the highest correlations between comfortable listening level and hearing loss appeared for 1.5 and 2 kHz (r = 0.73 and 0.77, respectively), followed by 1 kHz, 500 Hz and 3 kHz (all around 0.65). The correlation decreased to about 0.50 at either end (125 Hz and 6 kHz, respectively) and at 8 kHz the correlation was almost zero.

3.2 Sound quality ratings

The ratings of perceived sound quality were analysed by analysis of variance. This was done both for each individual subject (sources of variance: programs, systems, and replications; fixed model) and over all the subjects in the respective group (sources: programs, systems, subjects and replications; mixed model). The analyses were made separately for each rating scale.

3.2.1 Reliability

The intra-individual reliability of the ratings was studied by means of the "within cell mean square" (MSw) in the individual analyses, that is, the estimated variance of the replicated ratings (the sound quality ratings were made twice, in the second and in the fourth session). The smaller the variance of the repeated ratings, the better the reliability is. Significant *F* tests (5% level) for systems and systems x programs interaction were also used to indicate reliability.

Insert Table 2 about here

The results are given in Table 2 in terms of mean and range of MSw over subjects. MSw is generally lower for the NH subjects than for the HI subjects, that is, the former are more reliable. The value for the loudness scale is lower than for the other scales in both groups (0.87 for HI, 0.59 for NH), while the highest value appears for spaciousness (2.86 for HI, 2.00 for NH). There was no consistent tendency that a subject had similar MSw in all scales. The usual situation was rather that he/she had low MSw in some scale(s) and high in others. However, two of the NH subjects had low or fairly low MSw in all scales. The reason why the lowest values appear for loudness is probably that loudness is a more familiar dimension than the others; furthermore, the subjects had practiced loudness settings in the first session when determining the comfortable listening levels.

The F tests for programs, systems, and programs x systems in the individual analyses of variance all used MSw as the error term (fixed model). The number of subjects with significant F tests for the systems or programs x systems interaction is also given in Table 2. In the NH group almost all eight subjects had significant F tests in all scales. In the HI group at least eight of the twelve subjects had significant F tests in five scales. In the remaining three scales (loudness, clarity, total impression) the number was 4-6 subjects.

An estimate of inter-individual reliability (the agreement between subjects) is the r_b index (Winer, 1971, p. 283). Its maximum value is 1.00; the higher, the better. For the NH group the r_b values were between 0.79 and 0.93 (Table 2). For the HI group the corresponding values were about the same or somewhat lower, except for loudness and softness for which the r_b values were much lower. In the case of loudness this may be due to the adjustment to comfortable listening level, which decreases the variance due to programs and systems and thereby also decreases the value of r_b (Gabrielsson, 1979b). Although this argument applies to both groups, the range of comfortable loudness was smaller for the HI subjects than for the NH subjects (see Table 3 below).

All together the data indicated good reliability for the NH subjects. The reliability was lower for the HI subjects but acceptable for most of them; however, for some of the HI subjects it was partly unsatisfactory.

The NH listeners' ratings of the ideal values for the sound reproductions showed smaller MSw and thus higher reliability than the ratings of the real reproductions, see rightmost in Table 2. The corresponding data for the HI group are omitted for reasons discussed later in 3.2.3.

All MSw values reported here are higher than those obtained in earlier experiments on perceived sound quality of loudspeakers. In these experiments the subjects were well-trained listeners with long experience of high fidelity reproduction. For instance, in Gabrielsson & Lindström (1985) the average MSw values for the real reproductions varied between 0.62 to 0.89, and for ratings of the ideal values between 0.30 to 0.57.

3.2.2 Effects of reproduction systems

The results of the sound quality ratings for both subject groups are given in Table 3. The five reproduction systems appear in the columns and the six programs in the rows of each matrix. The values in the matrix are the arithmetic means over subjects. The row at the bottom of each matrix ("Mean") gives the means for each system in average over all programs. These means are also shown in the graphs of Figure 5. The results of the ratings of ideal values are

Pairwise comparisons between the systems may be made by means of the critical differences given in Table 4. There are many statistical tests for making such comparisons (Winer, 1971: Gabrielsson, 1979b). The two alternatives given here are a common t test (two-tailed) and the more conservative Tukey's HSD test. If the difference between any two systems exceeds the corresponding critical difference, it is statistically significant (5% level). For instance, the difference between systems No. 1 and No. 5 in fullness for the HI group is 1.2 (6.2-5.0, see the "Mean" row for fullness in Table 3). This difference is larger than either critical difference in fullness for this group (0.69 for the t test alternative, 0.99 for HSD, see Table 4) and thus considered significant.

From the evidence in Tables 3-4 and Figures 4-5 we can draw the following conclusions regarding the effects of the filtered reproductions on the perceived sound quality in this experiment:

(a) Fullness. Fullness is higher for systems with flat response below 1 kHz (systems Nos. 1 and 2) than for systems with successively reduced response below 1 kHz (systems Nos. 3, 4, and 5); the more reduction, the less fullness (compare systems Nos. 3 and 4). The greatest reduction below 1 kHz occurs for system No. 4 (cf. Figure 4), which also is rated least in fullness.

The significant systems x programs interaction for the HI group essentially means that some of the tendencies described here were more pronounced for certain programs (e.g. program No. 5) but less evident in others (program No. 2 and especially program No. 6).

(b) Loudness. Since the subjects themselves determined a comfortable listening level, it was expected that there should be no pronounced differences between the systems in loudness. This was confirmed for the HI group, while there are certain differences for the NH group. As suggested earlier, the range of comfortable loudness may be somewhat larger for the NH group. Note also that the definition of 5.0 in the loudness scale is "midway", which is not the same as "comfortable".

(c) Brightness. Brightness increases with increased level at higher frequencies (compare systems Nos. 1 and 2 between themselves and likewise systems Nos. 4 and 5) and with reduced response at lower frequencies (compare systems Nos. 2, 3, and 4). Highest brightness occurs for system No. 4, least brightness for system No. 1 (flat).

The significant systems x programs interaction for both groups means that there are some minor exceptions from the above tendencies. It is thus notable that the effects of the different filters are much smaller for program No. 6 than for the other programs. The spectrum of this program (a female singer) falls rapidly below 400 Hz, which means that this program sounds brighter than the other programs given for the NH group; they are discussed separately later.

Insert Table 3 and Figure 5 about here

Most values in Table 3 lie within 3.0-7.5 (the extreme values are 2.6 and 7.8). As seen in the "Mean" rows, there are sizable differences between the reproduction systems in almost all matrices. An obvious exception is in loudness for the HI group, where the means are practically identical (4.9-5.2). In the NH group, however, there are differences in loudness between the systems, although the NH subjects also made adjustments to comfortable listening level. It seems that the (subjective) range of comfortable loudness was somewhat larger for the NH subjects than for the HI subjects.

Further comparison between the two groups reveals (Figure 5) that the mean ratings as a rule lie somewhat higher for the HI subjects than for the NH subjects. With regard to the effects of the different systems the general tendencies are similar in both groups, but the NH listeners distinguish systems Nos. 3-5 from systems Nos. 1-2 more clearly than the HI listeners.

Analysis of variance was performed on the complete set of ratings in each scale for each of the two groups. The sources of variance were systems, programs, subjects, and interactions between these factors. The F tests were made according to a mixed model, i.e. the subjects considered as a random variable. The differences between the systems were thus tested by using the mean square for the system x subjects interaction as error term. The analyses are summarized in Table 4.

Insert Table 4 about here

There were significant differences between the systems in all scales for the NH group and in five of the eight scales for the HI group (the exceptions were loudness, softness, and spaciousness). The significant differences between the programs reflect obvious dissimilarities in frequency range, sound levels etc. of the programs and are as a rule not further discussed here. There were a few significant interactions between systems and programs. Furthermore, there were significant systems x subjects and programs x subjects interactions in all scales for both groups; this is a common finding in experiments on sound quality ratings (Gabrielsson & Lindström, 1985).

already from the beginning (cf. the brightness rating for program No. 6 at system No. 1 with the ratings for the other programs at the same system). It also means that the reductions of the lower frequencies have less effects for this program.

(e) Softness. Although there were no significant differences between systems for the HI group, the tendencies for this group are similar to those for the NH group. Softness is highest for system No. 1 (flat) and is reduced by increased response at higher frequencies (compare systems Nos. 1 and 2 between themselves and likewise systems Nos. 4 and 5) and especially by reduced response at lower frequencies as in systems Nos. 3, 4, and 5. The least soft or, in other words, the sharpest reproduction thus occurs for system No. 4.

e) Nearness. The systems with flat response below 1 kHz (Nos. 1 and 2) get the highest values, along with system No. 3 in the HI group. Increased response at higher frequencies seems to add somewhat to nearness (compare Nos. 1 and 2 between themselves). With 6 dB/octave reduction below 1 kHz and 6 dB/octave increase above 1 kHz as in system No. 3, nearness obviously decreases for the NH group (compare systems Nos. 2 and 3), while this effect is not found in the HI group. A closer inspection of the HI data shows that systems Nos. 2 and 3 are equivalent at the speech programs (Nos. 1 to 4), but that system No. 3 is inferior at the music programs (Nos. 5 to 6). With still further reduction of the lower frequencies, as in systems Nos. 4 and 5, nearness decreases for both groups, especially at system No. 4.

(f) Spaciousness. The ratings in this scale were the least reliable among all scales, and there was no significant difference among the systems for the HI group. For the NH group increased response at higher frequencies adds to spaciousness, when the response below 1 kHz is flat (compare systems Nos. 1 and 2). This also holds for the HI group. Furthermore, system No. 4 with its heavy bass reduction and emphasis on the treble is inferior in spaciousness to the others. However, this effect is not found in the HI group, for which systems Nos. 2-5 seem to be about equivalent.

(g) Clarity. In clarity there are also some differences between the subject groups. In both groups clarity increases when flat response below 1 kHz is combined with increased response at higher frequencies (compare systems Nos. 1 and 2). System No. 2 is best in clarity for the NH group. It is also best for the HI group with regard to the music programs (programs Nos. 5 and 6), while system No. 3 is the best for the speech programs (programs Nos. 1-4). There is thus an important interaction between systems and programs in the HI group.

The advantage with rising frequency response at higher frequencies is lost for the NH listeners, when it is combined with a corresponding 6 dB/octave reduction below 1 kHz as in system No. 3, which is equivalent to system No. 1 in clarity. System No. 5 is also equivalent to No. 1; in this case there is a heavy reduction at lower frequencies but a compensation in terms of a flat 6 dB higher level than for system No. 1 above 1 kHz. However, a still heavier reduction below 1 kHz, as in system No. 4, is not enough compensated for by a 6 dB/octave rise above 1 kHz. System No. 4 is thus the worst in clarity for the NH group.

For the HI listeners the situation is different. As noted above, systems No. 2 and 3 are the best for the music and the speech programs, respectively, closely followed by system No. 5. The lowest mean rating occurs for system No. 1, not for system No. 4 as in the NH group. However, there is an interaction: in fact system No. 4 is decidedly better than system No. 1 only for program No. 4 (the sentences in noise) and possibly at program No. 6. For the remaining four programs the two systems are about equivalent.

(h) Total impression. For the NH group system No. 2 with its combination of flat response below 1 kHz followed by increased response thereafter is the best, followed by system No. 1 with its flat response throughout. Systems Nos. 3-5 with cuts at lower frequencies and increases at higher frequencies are worse, especially system No. 4 with the heaviest cut and most steeply rising frequency response.

For the HI group system No. 4 is again the worst (however, not at program No. 4, the sentences in noise). System No. 2 is better than system No. 1, as in the NH group, and is best for the "realistic" speech programs (programs Nos. 1-2 with fluent speech) and for the music programs. However, for the PB-words and the sentences heard in noise systems Nos. 3 and 5 are equally good or better (note that there was an indication of systems x programs interaction for the HI group).

3.2.3 Ideal values

The listeners were asked to rate "how an ideal reproduction of each program should be" (see instructions in Appendix 1). This makes it possible to compare the ratings of the real reproductions with those for the (imagined) ideal reproductions. The idea has been successfully tried for ratings of loudspeakers by normal hearing listeners (e.g., Gabrielsson & Lindström, 1985). It should be noted that the ideal value is not necessarily the same as the maximum value; for instance, maximum clarity (10 in the clarity scale) may be felt as unnatural and exaggerated clarity. Usually the ideal values lie toward the upper end of the scale for clarity, fullness, and spaciousness, and roughly in the middle part of the scale for brightness and softness. Their exact positions on the scales vary, of course, with different programs.

It was suspected that the HI listeners might have difficulties in making this kind of ratings, especially listeners with hearing loss from early age. This suspicion was confirmed for many of the listeners by the experimenter's observations and follow-up questions. Furthermore, the data for the HI group showed practically no differences in rated ideal values among the various programs in any scale. For these reasons ideal values for the HI group are not included here.

For the NH group the results were similar to those in earlier experiments, and there were statistically significant differences between the ideal values for different programs in all scales but brightness. The mean ratings over all NH listeners are included in Table 3, and it is then easy to compare the ratings of the real reproductions with the ideal reproduction. For instance, the ideal value for program 1 in fullness is 7.1. Systems Nos. 1 and 2 are relatively close (6.7 and 6.4), while systems Nos. 3, 4, and 5 sound far too thin. The situation is partly similar for the other programs, but in program No. 5 (jazz music) all systems are far from the ideal value. Both music programs (Nos. 5 and 6) require more fullness than the speech programs.

In brightness system No. 2 (flat, +6) comes closest to the ideal, while system No. 1 sounds too dull (except for program No. 6). On the other hand systems Nos. 3, 4, and 5 with reduced lower frequencies and increased higher frequencies sound too bright. They also sound too sharp (not soft enough) as seen from the values in the softness scale. The program requiring most softness is the female voice reading a fairy tale (program No. 1). In nearness the ideal values are higher for the speech programs (Nos. 1-4) than for the music programs. All reproductions fall short of the ideal values (sound too distant) except systems Nos. 1 and 2 for program No. 6. The music programs (Nos. 5 and 6) require more spaciousness than the speech programs. All reproductions sound too narrow (not enough open, spacious) in relation to the ideal value. Also in clarity there is in most cases a considerable negative deviation from the ideal value, that is, no reproduction of any program reaches the desired clarity.

3.2.4 Relations between perceptual scales and total impression

The product moment correlation between each single perceptual scale and the total impression scale is given for each program and listener group in Table 5. For instance, the correlation between fullness and total impression for the HI group at program No. 1 was .42.

Insert Table 5 about here

As a rule the correlations are clearly higher for the NH group than for the HI group. An exception from this occurs for clarity at program No. 3 (the PB words).

For the HI group the highest correlations with total impression appear for clarity (.49 to .70). Then follow spaciousness (.49 to .57), fullness (.36 to .61), nearness (.30 to .57), softness (.16 to .58), loudness (-.13 to .46), and brightness (-.06 to .19). In the NH group the order is about the same, but fullness comes first (.68 to .83), followed by clarity (.50 to .83), nearness (.63 to .81), spaciousness (.53 to .74), softness (.52 to .69), loudness (.25 to .52), and brightness (-.34 to -.48).

The fact that brightness correlates negatively with total impression for the NH group is mainly due to the characteristics of systems Nos. 3, 4, and 5. These systems with reduced response at lower frequencies and increased response at higher frequencies sound brighter than systems Nos. 1 and 2 but are clearly inferior to systems Nos. 1 and 2 in total impression (cf. Table 3). For the HI group systems Nos. 3, 4, and 5 likewise sound brighter, but they are about equivalent to systems Nos. 1 and 2 in total impression. The correlation between brightness and total impression for the HI group thus approaches zero.

The multiple correlation of all perceptual scales with total impression was high for the NH group (.83 to .94, see Table 5), lower but still relatively high for the HI group (.69 to .83). Usually only two or three scales were required to almost reach the maximum multiple correlation; in most cases these included clarity, and/or fullness, and/or spaciousness.

The results for the NH group are similar to those obtained for NH listeners judging loudspeaker reproductions (Gabrielsson & Lindström 1985), except for the negative correlation between brightness and total impression.

The data above refer to the <u>groups</u> of HI and NH listeners. If the correlations between perceptual scales and total impression are studied separately for each <u>individual</u>, it turns out that the multiple correlation is <u>as high</u> for individual HI listeners (varying between them from .65 to .96, median value .85) as it is for individual NH listeners (.72 to .94, median value .88). This suggests that the lower correlations for the HI listeners as a group are due to a more inter-individually varying structure of the correlations between perceptual scales and total impression in the HI group than in the NH group. This may be expected from the larger variability in hearing ability among the HI listeners than among the NH listeners. This was confirmed by a factor analysis which showed that all NH listeners but one had a similar structure of the correlations, while there were three subgroups with different correlation structures among the HI listeners.

3.3 Measures of speech intelligibility

Speech intelligibility was measured by two methods: PB words (only for the HI listeners) and sentences heard in noise.

The percentage correctly recognized PB words was practically identical for all five systems (82-85%) in average over all HI listeners. However, there were individual differences and indications of interactions between systems and individuals. Half of the HI listeners had 90% or more correct recognition and showed no or small differences in recognition among the different systems; these subjects as a rule had the best hearing (cf. the audiograms for listeners Nos. 3, 4, 5, 6, 8, and 12, Table 1). For five listeners with average recognition of 67-83% and more pronounced hearing loss (subjects Nos. 1, 2, 7, 10 and 11) the recognition was worse for one or more of the five systems (subject No. 1 excepted). For the listener with the worst hearing impairment (subject No. 9) the average recognition was about 50%, but there was a large variation around this average: 22% for system No. 5, 52-58% for systems Nos. 1-3, and 70% for system No. 4. In general then, the more pronounced hearing loss, the worse recognition, and the more differences in recognition among the five systems.

Speech intelligibility was also measured by the signal-to-noise ratio for correct recognition of sentences heard in noise. For the NH group the individual S/N thresholds, in average over all systems, varied from -6.0 to -7.8 dB, but there was practically no difference among the S/N thresholds for the different systems in average over all listeners (it varied from -7.1 to -7.3 dB). Within single listeners the range of the S/N threshold for different systems varied from 0.6 to 3.0 dB.

For the HI listeners the individual S/N thresholds, in average over all systems, varied very much, from -6.4 to +10.7 dB, median value -2.7 dB. The ratio was negative for nine out of the twelve listeners, and positive for the remaining three (these values were 0.9, 3.0, and 10.7, respectively). Analysis of variance (systems and listeners as sources of variance, the residual being the error term) indicated significant differences (5% level; $F=2.63$, df 4/44) between the systems. The actual difference is between the S/N threshold for system No. 1, -0.2 dB, and the S/N threshold for the remaining four systems, which was practically identical for all of them, -1.5 to -1.8 dB.

There was an obvious correlation between hearing loss and S/N threshold: the less hearing loss, the better (that is, the more negative) S/N threshold. For all but two listeners system No. 1 (flat) gave the worst or next to worst S/N threshold. The difference in S/N threshold between the remaining four systems, Nos. 2-5, varied from 0.4 to 2.3 dB among the different listeners (except for subject No. 10), and which system was the best varied for different listeners.

3.4 Relations between sound quality dimensions and intelligibility measures

Since the PB words did not differentiate between the five systems, the remaining measure of speech intelligibility for further comparison is the S/N threshold.

This comparison was limited to program No. 4 (sentences in noise). With regard to the NH listeners, there was practically no difference in S/N threshold among the five systems. Consequently the correlations between different sound quality scales and S/N threshold were close to zero. This holds for the NH listeners considered as a group. However, for individual NH listeners there were examples of substantial correlations (positive or negative) between some quality scale(s) and the S/N threshold.

For the HI group there were some substantial correlations between quality scales and S/N threshold, see Table 6 (the correlation is based on 60 cases, 12 subjects x 5 systems).

Insert Table 6 about here

The highest correlations appear for clarity, loudness (both -.62), and spaciousness (-.47). In other words, the more clear, loud, and spacious reproduction, the better recognition of speech. (The correlations are negative, since a more negative S/N threshold means better recognition.) The essential difference in S/N threshold among the systems was that system No. 1 (flat) was inferior to the others. By checking the data for program No. 4 in Table 3 it can be seen that system No. 1 was also inferior to the others in clarity and spaciousness, as well as in total impression.

With regard to individual HI listeners, the correlations showed a more varying pattern. A factor analysis indicated that the twelve listeners could be subdivided into at least four categories with different characteristics regarding the correlations between perceptual scales and the S/N threshold.

3.5 Answers to the questionnaire

Most subjects in both groups said that they understood the meaning of the rating scales without difficulty. A few subjects hesitated about one or two scales, mostly about spaciousness. This scale was also the least reliable (Table 2). Despite this there were clear differences between the systems in this scale as well (Tables 3-4), and the correlation between spaciousness and total impression was rather high (Table 5).

The listeners were also asked to rate the importance of the different scales for good sound quality. The rating scale was graded in five steps: 5 = very important, 4 = rather important, 3 = midway, 2 = not especially important, 1 = not important at all. The ratings were made separately for each of the six programs. It was apparent that the results differed between the speech programs (Nos. 1-4) and the music programs (Nos. 5-6). The mean ratings over all subjects in the respective group are given in Table 7.

Insert Table 7 about here

In the HI group all scales seem to be important (the lowest mean value is 3.8). For speech the most important scales are clarity, nearness, and loudness. For music there is less distinction among the scales. It can be noted that softness and spaciousness have gained in importance, while clarity and nearness have decreased. Regarding the importance of softness some subjects remarked that the female singer (program No. 6) sometimes sounded very shrill.

In the NH group there is a clearer distinction among the scales, but the tendencies are similar to those in the HI group. Clarity, loudness, and nearness are most important for speech, while fullness and spaciousness are least important. For music, however, fullness and spaciousness are the most important, and brightness as well as softness have gained in importance. Clarity is still important but not as much as for speech, and nearness is now the least important scale. Of course, these results partly depend on the type of music programs used here. With other programs the relative importance of the scales may be another.

Practically all subjects thought that the task of making sound quality ratings became successively easier for them during the progress of the experiment. The experiment as a whole was considered as demanding but interesting. Some of the HI listeners thought, not unexpectedly, that listening to sentences in noise was very tiring.

4 DISCUSSION

4.1 Ratings of sound quality

The reliability of the sound quality ratings was good for the NH group and satisfactory for the HI group. With regard to individual subjects the reliability was partly unsatisfactory for some HI listeners. The effects of the different frequency responses were obvious in all rating scales for the NH group and in most scales for the HI group. This also attests to the validity of the scales. The validity is further confirmed by the fact that all perceptual scales (clarity, fullness etc.) were correlated, to varying degrees, with the scale for total impression. The multiple correlation of all or some of the perceptual scales with total impression varied between .83 and .94 for the NH group, and between .73 and .83 for the HI group (Table 5). Subjectively all perceptual scales were considered to be more or less important for good sound reproduction (Table 7). Clarity, loudness, and nearness were judged most important for reproduction of speech. The other scales were relatively more important for the reproduction of music.

As seen in Table 3, none of the reproductions was considered as especially good. System No. 2, which was best in total impression, got a mean rating over programs close to 6.0 in both groups, that is between "midway" and "rather good". However, there were numerous ratings higher than 7.0 in the various scales for different program x system combinations. With regard to the rating accuracy it can be noted that all NH subjects and 9 out of 12 HI subjects used decimals in their ratings.

The effects of the different frequency responses may be conveniently described using the flat response of system No. 1 as a reference. The results can then be summarized as follows:

A 6 dB/octave increase at higher frequencies combined with flat response at lower frequencies (system No. 2) resulted in increased brightness, nearness, spaciousness, clarity, and better total impression; softness decreased, and fullness remained about the same. This holds for both listener groups.

When a 6 dB/octave rise above 1 kHz was combined with a 6 dB/octave decrease below 1 kHz (system No. 3), fullness and softness decreased, while brightness increased in comparison with the flat response in system No. 1. In the remaining scales the effects were different for the two listener groups. In the NH group spaciousness and clarity remained the same, while nearness decreased as well as the rating of the total impression. However, in the HI group spaciousness and clarity increased (although the data for spaciousness are somewhat unreliable), nearness remained the same, and the rating of the total impression also remained the same or possibly improved a little.

With the 6 dB/octave increase above 1 kHz combined with 12 dB/octave reduction below 1 kHz (system No. 4) fullness, softness, and nearness decreased, and brightness increased in both groups in comparison with system No. 1. These effects are among the largest in the whole set of data. In the NH group spaciousness and clarity also deteriorated, as well as the total impression. In the HI group spaciousness and clarity remained about the same, while there was a certain deterioration in total impression.

With essentially the same type of 12 dB/octave reduction at lower frequencies and a flat, but 6 dB higher, response above 1 kHz (system No. 5), fullness, softness, and nearness decreased, and brightness increased in comparison with system No. 1 in the NH group. The tendencies were the same but weaker in the HI group. Spaciousness and clarity remained the same in the NH group but increased somewhat in the HI group. The total impression got worse for the NH group, but remained about the same for the HI group.

The psychophysical relations suggested by all these data agree fairly well with those discussed by Gabrielsson & Sjögren (1979a, 1979b). Fullness depends very much upon the contribution by lower frequencies, and brightness upon the contribution by higher frequencies. Softness also depends much on the lower frequencies, and a very steeply rising frequency response as in system No. 4 tends to give sharpness. Clarity and nearness are favored by a certain emphasis on higher frequencies, especially for the HI listeners, but not to such an extreme degree as in system No. 4.

The differences in ratings between the two listener groups essentially referred to the systems with reductions at lower frequencies (systems Nos. 3-5). These systems were considered as thinner and brighter by both groups. The NH group also rated them as sharper, more distant, less spacious, less clear and worse in total impression than either or both of systems Nos. 1 and 2; system No. 4 was definitely worst. However, in the HI group systems Nos. 3 and 5 came relatively close to systems Nos. 1 and 2 in these scales; system No. 4 was worst but not so far behind.

The ratings of ideal values worked well with the NH subjects and provided valuable information for the interpretation of the results. Although the procedure was not enough successful for the HI subjects (we estimated that about half of the HI listeners were able to follow the instruction), it should be possible to design a proper procedure (e.g., by means of more detailed instructions and explanations) for hearing impaired listeners as well. Information about ideal values in various scales could give important knowledge for judgment and selection of hearing aids.

4.2 Sound quality and intelligibility

For the NH group there were obvious differences among the systems in all sound quality scales but no difference with regard to the S/N threshold. For the HI group there were also rather obvious differences among the systems in most quality scales, but there was no distinction among them for the PB words and only a limited distinction (system No. 1 versus the others) in the S/N threshold. Thus the quality ratings provided better differentiation among the systems than the two intelligibility measures.

These statements apply to the NH and HI listeners considered as groups. Further inspection of the data for individual HI listeners shows that the effects of the systems were similar on the PB-word score and on the S/N treshold for half of them. However, for three of these listeners neither of those methods differentiated among the systems. With regard to the relation between S/N thresholds and sound quality scales there were also considerable inter-individual variations. However, most HI listeners showed a substantial (negative) correlation between the S/N threshold and the clarity, loudness, and spaciousness scales (cf. 3.4).

As measured in this experiment, there is thus no simple relationship neither between the two intelligibility measures (cf. also Hagerman, 1984b), nor between these and the sound quality scales. The only clear-cut distinction among the systems from the intelligibility measures was that system No. 1 got worse S/N threshold than the other systems in the HI group. This system was also worst in clarity, spaciousness, and total impression for the program with sentences in noise (cf Table 3, data for program No. 4). In this limited respect there was thus a certain correspondence between the intelligibility and sound quality measures. However, the sound quality data in Table 3 generally provide much more information and show that there were numerous other distinctions among the systems as well. Our results do not support the idea of a general conflict between intelligibility and sound quality measures. The conclusion is rather that the relationship between intelligibility and sound quality varies depending on many factors, such as which measures of intelligibility and sound quality are used, the type of programs, the specific hearing loss, and the persons's listening habits.

The ratings of sound quality were more time-consuming because of the many scales and the repeated ratings. In clinical practice time limitations necessitate a reduction of the number of scales and ratings. The minimum would be an overall quality scale, corresponding to total impression here (but not necessarily with that name). This may be enough for choosing the "best" hearing aid among those under consideration, but repeated ratings should be made in order to check the reliability of the ratings. One or more other scales may be included to give more detailed information, e.g. to differentiate among aids which get about the same overall quality rating. The choice of programs is important. They should preferably be representative for the individual's own situation and listening habits, that is, what he needs and wants to hear (Hagerman & Gabrielsson, 1985).

Among the systems used in this investigation system No. 2 was judged as generally best in sound quality by the NH group and best for all "realistic" programs (programs Nos. 1, 2, 5, and 6) by the HI group. For the two "artificial" programs, the PB words and the sentences in noise, systems Nos. 2, 3, and 5 were about equivalent in the HI group. System No. 2 meant a flat response below 1 kHz and a 6 dB/octave rise at higher frequencies. This points to the importance of an extended low frequency region as well as a certain emphasis on higher frequencies. This is reminiscent of the results found by Punch and his co-workers as well as by Franks (cf. the introduction). In the HI group the systems with reductions at lower frequencies (systems Nos. 3-5) were less inferior to system No. 2 than in the NH group. This may indicate that masking effects due to lower frequencies are more detrimental to HI listeners (these effects are reduced in systems Nos. 3-5). This is probably more critical for speech than for music. Franks' (1982) proposal that hearing aids should have a switching mechanism allowing to include low frequencies when listening to music and exclude them when listening to speech seems reasonable.

A further analysis of the relationships between the different frequency responses and the perceptual dimensions would require a more detailed physical analysis including the spectral contents of the different programs and should also include the audiometric data of the individual HI subjects (preferably also other relevant data). It is possible that such an analysis would clarify various interactions between programs and systems, as well as individual differences, which have not been discussed or only briefly hinted at in this report.

5 ACKNOWLEDGMENTS

We wish to express our gratitude to the staff of the hearing clinic at the Rehabilitation Institute for Hard of Hearing, Deaf and Deaf-blind (AMI-s hörsel), Uppsala, for their kind cooperation, to Sven Jalmell for electronic constructions, to Bertil Johansson for valuable discussions, and to our subjects for their patient work.

6 REFERENCES

FRANKS, J.R. (1982). Judgments of hearing aid processed music. Ear and Hearing, 3, 18-23.

GABRIELSSON, A. (1979a). Dimension analyses of perceived sound quality of sound-reproducing systems. Scandinavian Journal of Psychology, 20, 159-169.

GABRIELSSON, A. (1979b). Statistical treatment of data from listening tests on sound-reproducing systems. Technical Audiology Reports No. 92. Stockholm: Karolinska Institute.

GABRIELSSON, A. (1981). Problems and methods in judgments of perceived sound quality. Technical Audiology Reports No. 103. Stockholm: Karolinska Institute; Abstract in Journal of the Acoustical Society of America, 70, Suppl. 1, S64.

GABRIELSSON, A., & LINDSTRÖM, B. (1981). Scaling of perceptual dimensions in sound reproduction. Technical Audiology Reports No. 102. Stockholm: Karolinska Institute.

GABRIELSSON, A., & LINDSTRÖM, B. (1985). Perceived sound quality of high-fidelity loudspeakers. Journal of the Audio Engineering Society, 33, 33-53.

GABRIELSSON, A., & SJÖGREN, H. (1979a). Perceived sound quality of hearing aids. Scandinavian Audiology, 8, 159-169.

GABRIELSSON, A., & SJÖGREN, H. (1979b). Perceived sound quality of sound-reproducing systems. Journal of the Acoustical Society of America, 65, 1019-1033.

GABRIELSSON, A., FRYKHOLM, S.Å., & LINDSTRÖM, B. (1979). Assessment of perceived sound quality in high fidelity sound-reproducing systems. Technical Audiology Reports No. 93. Stockholm: Karolinska Institute.

GABRIELSSON, A., LINDSTRÖM, B., & ELGER, G. (1983). Assessment of perceived sound quality of eighteen high fidelity loudspeakers. Technical Audiology Reports No. 106. Stockholm: Karolinska Institute.

GABRIELSSON, A., HAGERMAN, B., BERG, C., OVEGÅRD, A., & ÄNGGÅRD, L. (1980). Clinical assessment of perceived sound quality in hearing aids. Technical Audiology Reports No. 98. Stockholm: Karolinska Institute.

HAGERMAN, B. (1982). Sentences for testing speech intelligibility in noise. Scandinavian Audiology, 11, 79-87.

HAGERMAN, B. (1984a). Some aspects of methodology in speech audiometry. Scandinavian Audiology, Supplementum 21.

HAGERMAN, B. (1984b). Clinical measurement of speech reception threshold in noise. Scandinavian Audiology, 13, 57-63.

HAGERMAN, B., & GABRIELSSON, A. (1985). Questionnaires on desirable properties of hearing aids. Scandinavian Audiology, 14, 109-111.

HARFORD, E.R., & FOX, J. (1978). The use of high-pass amplification for broad-frequency sensorineural hearing loss. Audiology, 17, 10-26.

HARRIS, R.W., & GOLDSTEIN, D.P. (1979). Effects of room reverberation upon hearing aid quality judgments. Audiology, 18, 253-262.

HARRIS, R.W., & GOLDSTEIN, D.P. (1985). Hearing aid quality judgments in reverberant and nonreverberant environments using a magnitude estimation procedure. Audiology, 24, 32-43.

JERLVALL, L., ALMQVIST, B., OVEGÅRD, A., & ARLINGER, S. (1983) Clinical trial of in-the-ear hearing aids. Scandinavian Audiology, 12, 63-70.

LEVITT, H. (1978). Methods for the evaluation of hearing aids. In C. Ludvigsen & J. Barfod (Eds.), Sensorineural hearing impairment and hearing aids, Scandinavian Audiology, Supplementum 6, 199-240.

PUNCH, J.L. (1978). Quality judgments of hearing-aid-processed speech and music by normal and otopathologic listeners. Journal of the American Audiology Society, 3, 179-188.

PUNCH, J.L., & BECK, E.L. (1980). Low-frequency response of hearing aids and judgments of aided speech quality. Journal of Speech and Hearing Disorders, 45, 325-335.

PUNCH, J.L. & PARKER, C.A. (1981). Pairwise listener preferences in hearing aid evaluation. Journal of Speech and Hearing Research, 24, 366-374.

PUNCH, J.L., MONTGOMERY, A.A., SCHWARTZ, D.M., WALDEN, B.E., PROSEK, R.A., & HOWARD, M.T. (1980). Multidimensional scaling of quality judgments of speech signals processed by hearing aids. Journal of the Acoustical Society of America, 68, 458-466.

THOMPSON, G., & LASSMAN, F. (1969). Relationship of auditory distortion test results to speech discrimination through flat vs. selective amplifying systems. Journal of Speech and Hearing Research, 12, 594-606.

THOMPSON, G., & LASSMAN, F. (1970). Listener preference for selective vs. flat amplification for a high-frequency hearing-loss population. Journal of Speech and Hearing Research, 13, 670-672.

WINER, B.J. (1971). Statistical Principles in Experimental Design (2nd ed.) New York: McGraw Hill.

Subj. No.	.125	.25	.5	1	1.5	2	3	4	6	8	Type
	Frequency (kHz)										
1	17	05	25	55	65	58	54	70	81	85	S
2	35	45	75	80	80	75	70	55	70	45	S
3	10	10	30	40	45	50	45	60	65	65	S
4	25	20	15	15	30	50	60	60	95	75	S
5	30	35	40	45	65	75	85	80	90	70	S
6	20	25	40	55	55	50	75	80	75	70	S
7	40	50	55	65	80	75	85	95	95	85	S
8	20	30	40	45	55	50	50	50	45	45	S
9	60	65	85	80	85	85	100	115	>120	>100	M
10	40	75	80	90	95	95	90	90	90	80	M
11	15	30	45	65	70	75	85	90	>120	>100	S
12	25	25	25	30	35	25	35	60	80	>100	M

Table 1. Individual pure tone thresholds in dB HL and types of impairment for the hearing impaired group. S=sensorineural, M=mixed (conductive + sensorineural).

HEARING IMPAIRED (HI)

	Mean	MSw Range	Sign. F-test (max=12)	r_b
Fullness	1.97	1.24-3.25	10	0.87
Loudness	0.87	0.10-1.83	4	0.52
Brightness	1.79	0.64-3.33	9	0.93
Softness	1.61	0.55-2.30	8	0.47
Nearness	2.50	1.18-3.83	8	0.78
Spaciousness	2.86	1.64-4.70	8	0.74
Clarity	2.05	0.43-2.91	4	0.87
Total impr.	1.94	0.89-3.93	6	0.89

NORMAL HEARING (NH)

	Mean	MSw Range	Sign. F-test (max=8)	r_b	MSw ideal Mean
Fullness	1.62	0.50-2.72	8	0.90	1.10
Loudness	0.59	0.13-0.99	7	0.86	0.27
Brightness	1.34	0.50-2.47	8	0.93	0.96
Softness	1.55	0.25-2.21	7	0.88	1.35
Nearness	1.52	0.20-2.28	8	0.89	1.06
Spaciousness	2.00	0.44-3.70	6	0.79	0.95
Clarity	1.55	0.82-2.24	6	0.79	0.69
Total impr.	1.50	0.54-2.25	8	0.87	

Table 2. Mean value and range over subjects for the MSw index in the different rating scales, number of subjects with significant *F* tests in the respective scales, and value of the r_b index. Mean value of MSw for ratings of ideal values is given for the NH group.

Program	HEARING IMPAIRED System 1	2	3	4	5	NORMAL HEARING System 1	2	3	4	5	Ideal
					FULLNESS						
1	6.5	6.7	5.1	4.2	5.2	6.7	6.4	4.5	3.2	4.9	7.1
2	5.9	5.2	4.5	4.1	4.2	5.8	5.7	4.3	3.7	4.0	6.5
3	7.0	7.0	5.9	5.0	5.8	6.1	6.3	3.9	3.2	3.5	7.0
4	5.9	6.0	5.3	3.9	4.5	6.4	6.4	3.9	2.6	4.6	6.9
5	7.2	6.8	5.2	4.0	5.3	5.7	5.8	4.0	3.3	4.3	8.1
6	4.5	5.0	4.7	4.2	4.8	6.2	6.1	4.6	4.4	4.9	7.6
Mean	6.2	6.1	5.1	4.2	5.0	6.2	6.1	4.2	3.4	4.4	
					LOUDNESS						
1	5.3	5.2	5.6	5.1	5.1	4.7	5.5	4.7	4.3	4.9	5.1
2	5.1	5.1	4.8	4.2	4.5	4.5	5.3	4.9	3.6	4.0	5.9
3	5.2	5.3	5.2	4.9	5.0	4.7	5.3	4.9	4.2	4.1	5.3
4	4.6	4.7	5.0	4.7	4.4	4.2	4.7	4.1	3.5	3.9	5.2
5	5.2	5.4	5.5	5.5	5.2	4.5	5.0	5.1	4.2	4.2	6.4
6	4.8	5.1	5.0	5.2	5.1	5.1	5.9	5.7	6.0	5.2	5.5
Mean	5.0	5.1	5.2	4.9	4.9	4.6	5.3	4.9	4.3	4.4	
					BRIGHTNESS						
1	5.4	5.7	6.2	7.2	6.6	3.4	4.8	6.2	6.8	6.2	5.3
2	4.0	5.2	6.4	6.7	6.3	3.8	4.7	6.3	6.3	5.7	5.3
3	3.9	4.0	5.5	6.4	5.5	3.3	4.1	5.6	6.1	6.0	4.5
4	4.0	4.2	6.0	7.3	6.6	3.1	4.0	6.0	6.7	5.1	4.8
5	5.0	6.0	7.5	7.5	6.9	4.1	5.5	7.1	7.0	6.3	5.2
6	6.9	7.0	7.3	7.8	7.3	5.9	5.6	6.3	6.8	6.2	5.5
Mean	4.9	5.4	6.5	7.2	6.5	3.9	4.8	6.3	6.6	5.9	
					SOFTNESS						
1	6.2	5.9	5.6	5.0	5.9	6.8	6.1	4.9	4.0	4.6	7.5
2	5.8	5.4	5.9	5.3	5.3	6.4	5.7	4.3	4.2	4.8	6.0
3	6.2	6.1	5.7	5.1	5.6	7.0	5.9	4.1	3.7	5.4	6.4
4	6.2	5.9	6.0	5.7	6.0	6.9	6.4	4.6	3.8	5.3	6.4
5	6.6	5.8	4.9	4.3	5.5	6.7	5.3	3.8	3.8	4.7	6.1
6	5.3	5.3	5.2	4.7	5.6	5.8	5.3	4.4	3.7	4.8	6.7
Mean	6.1	5.7	5.6	5.0	5.7	6.6	5.8	4.4	3.9	4.9	

Table 3. Mean ratings in all perceptual scales for the different systems and programs. Mean ideal ratings are given for the NH group. (The table continues on the next page.)

Program	HEARING IMPAIRED System 1	2	3	4	5	NORMAL HEARING System 1	2	3	4	5	Ideal
NEARNESS											
1	7.1	6.7	6.8	5.6	5.9	6.5	7.0	5.3	3.8	5.5	7.4
2	5.5	5.7	5.6	4.0	4.4	5.2	5.8	4.7	3.1	3.9	7.1
3	6.5	7.0	6.8	5.2	6.4	6.3	6.7	4.6	3.4	3.6	7.1
4	5.3	5.7	5.7	4.7	5.0	5.1	5.7	4.0	3.0	4.3	7.1
5	6.2	7.1	5.9	5.3	5.8	5.1	5.3	4.7	3.1	3.5	5.8
6	4.9	5.5	4.9	4.5	5.3	5.2	5.1	4.4	4.3	3.9	5.2
Mean	5.9	6.3	6.0	4.9	5.5	5.6	5.9	4.6	3.5	4.1	
SPACIOUSNESS											
1	5.0	5.5	6.0	5.4	5.7	4.1	5.9	4.5	2.9	4.6	6.9
2	4.3	5.2	5.1	4.5	4.9	4.2	5.1	4.6	3.3	4.3	6.9
3	4.8	5.3	6.0	5.0	5.7	4.2	5.0	4.5	3.3	3.6	6.2
4	3.1	3.7	5.4	4.9	4.9	3.8	4.7	3.7	2.8	3.7	6.1
5	6.0	6.6	6.1	5.8	6.3	4.3	5.3	4.5	3.9	4.1	7.9
6	5.8	6.6	5.9	6.0	6.4	6.6	5.7	5.2	4.8	5.6	7.2
Mean	4.8	5.5	5.8	5.3	5.7	4.5	5.3	4.5	3.5	4.3	
CLARITY											
1	7.0	7.4	7.4	7.2	7.4	6.6	7.5	6.5	5.4	6.2	8.1
2	4.8	5.7	6.0	4.9	5.4	4.3	6.0	5.5	4.1	4.7	8.1
3	6.8	7.0	7.6	6.5	7.3	6.2	6.6	6.4	5.3	6.1	8.4
4	5.0	5.9	7.1	6.8	6.5	5.5	6.1	5.2	4.7	5.3	8.5
5	6.7	7.4	6.8	6.4	6.8	4.9	6.2	5.4	4.6	5.2	7.2
6	4.6	6.2	5.6	5.2	5.7	6.8	6.5	5.8	5.7	6.1	7.4
Mean	5.8	6.6	6.8	6.2	6.5	5.7	6.5	5.8	5.0	5.6	
TOTAL IMPRESSION											
1	6.6	7.5	6.5	5.7	6.8	6.2	7.0	5.1	3.5	4.9	
2	4.1	5.0	4.6	3.6	4.0	4.5	5.8	4.6	2.9	4.0	
3	6.3	6.7	6.9	5.7	6.6	5.8	6.2	4.6	3.9	4.5	
4	4.6	5.6	6.3	5.3	5.5	4.7	5.3	3.8	2.8	4.1	
5	6.9	7.3	6.4	5.2	6.0	4.6	5.5	4.4	3.5	4.1	
6	4.6	5.2	4.9	4.4	5.2	6.1	5.8	5.1	4.4	5.0	
Mean	5.5	6.2	5.9	5.0	5.7	5.3	5.9	4.6	3.5	4.4	

Table 3 continued.

		S	P	SxP	Critical diff. t test	Critical diff. HSD
Fullness	HI	*	*	*	0.69	0.99
	NH	*			0.68	0.97
Loudness	HI		(*)			
	NH	*	*		0.40	0.57
Brightness	HI	*	*	*	0.55	0.79
	NH	*	*	*	0.47	0.67
Softness	HI					
	NH	*			0.53	0.75
Nearness	HI	*	*		0.56	0.80
	NH	*	*	(*)	0.58	0.82
Spaciousness	HI		*	(*)		
	NH	*	*		0.74	1.05
Clarity	HI	*	*	*	0.47	0.67
	NH	*	*		0.61	0.86
Total impr.	HI	*	*	(*)	0.56	0.80
	NH	*	*		0.68	0.97

Table 4. Results from analyses of variance and significance tests. S=systems, P=programs, *=significant at .05 level, (*)=significant at .10 level. Further explanation in text.

		PROGRAM					
		1	2	3	4	5	6
Fullness	HI	.42	.38	.55	.36	.61	.58
	NH	.76	.72	.77	.68	.77	.83
Loudness	HI	.18	.16	.24	.39	-.13	.46
	NH	.48	.52	.47	.47	.32	.25
Brightness	HI	-.02	.03	.13	.19	-.06	.11
	NH	-.45	-.34	-.48	-.40	-.39	-.41
Softness	HI	.17	.31	.20	.16	.58	.22
	NH	.64	.52	.67	.62	.63	.69
Nearness	HI	.36	.30	.54	.50	.35	.57
	NH	.81	.63	.70	.64	.63	.64
Spaciousness	HI	.49	.55	.50	.51	.51	.57
	NH	.69	.53	.53	.57	.74	.71
Clarity	HI	.54	.49	.70	.69	.65	.54
	NH	.83	.67	.50	.72	.75	.73
Multiple correlation	HI	.73	.69	.77	.73	.83	.80
	NH	.94	.83	.84	.87	.90	.90

Table 5. Correlations between perceptual scales and total impression. HI=hearing impaired, NH=normal hearing.

Fullness	-.35
Loudness	-.62
Brigthness	-.15
Softness	-.02
Nearness	-.37
Spaciousness	-.47
Clarity	-.62
Total impr.	-.32

Table 6. Correlation between S/N treshold and perceptual scales for the HI group at program No. 4.

	HI group		NH group	
	Speech	Music	Speech	Music
Clarity	4.7	4.1	4.6	4.2
Fullness	3.9	4.0	2.8	4.3
Brightness	3.9	4.0	3.3	4.1
Softness	4.0	4.3	3.5	3.9
Spaciousness	3.9	4.2	2.6	4.5
Nearness	4.5	3.8	3.6	2.9
Loudness	4.5	4.3	4.0	4.0

Table 7. Ratings of the importance of various perceptual dimensions for the reproduction of speech and music.

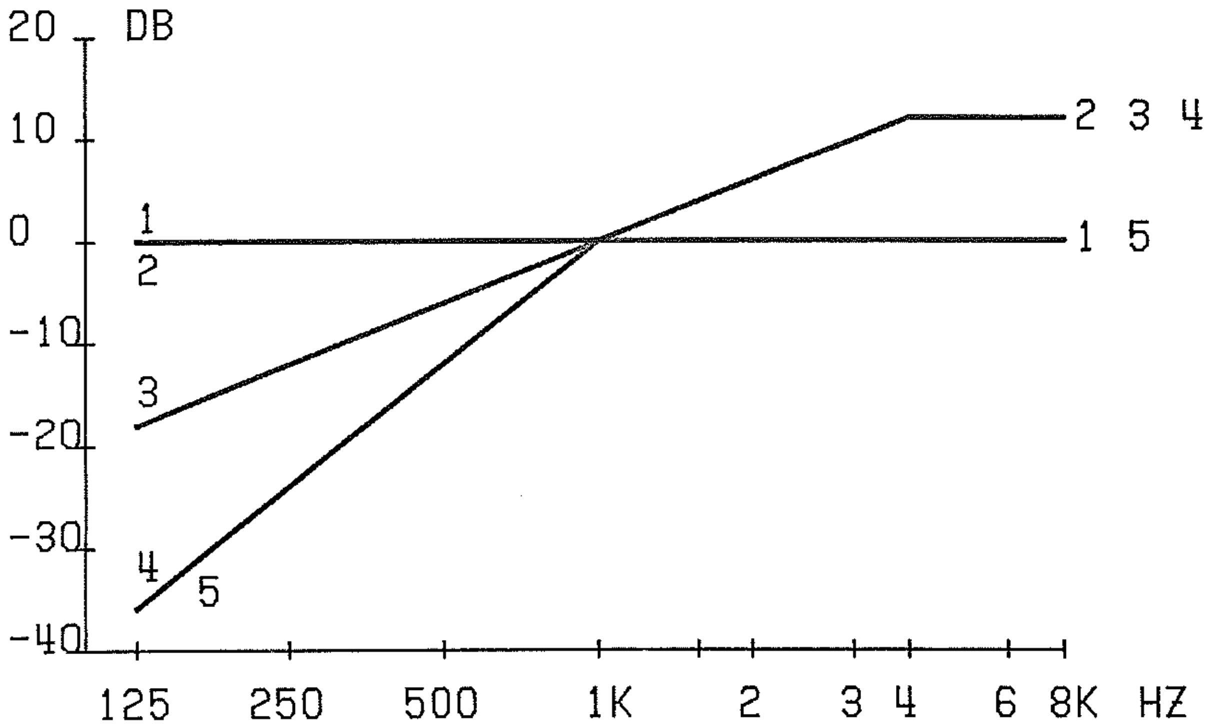

Figure 1. Schematic electrical responses of the five filters.

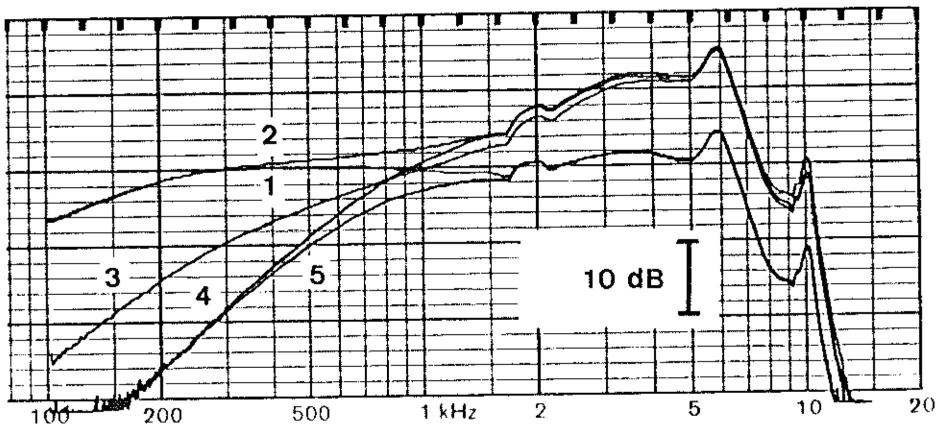

Figure 2. Acoustic frequency responses of the five reproduction systems.

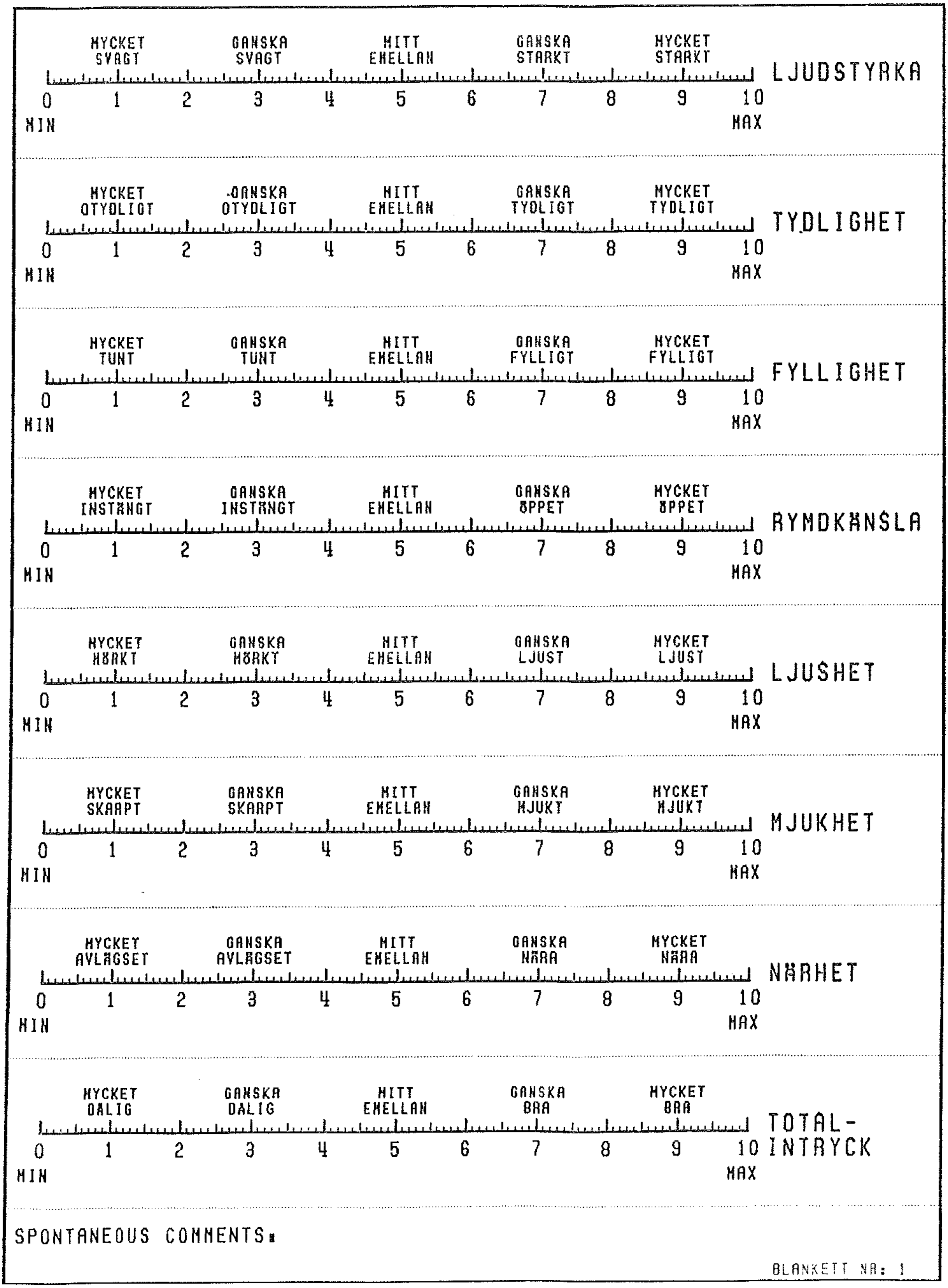

Figure 3a. Example of the response form (Swedish).

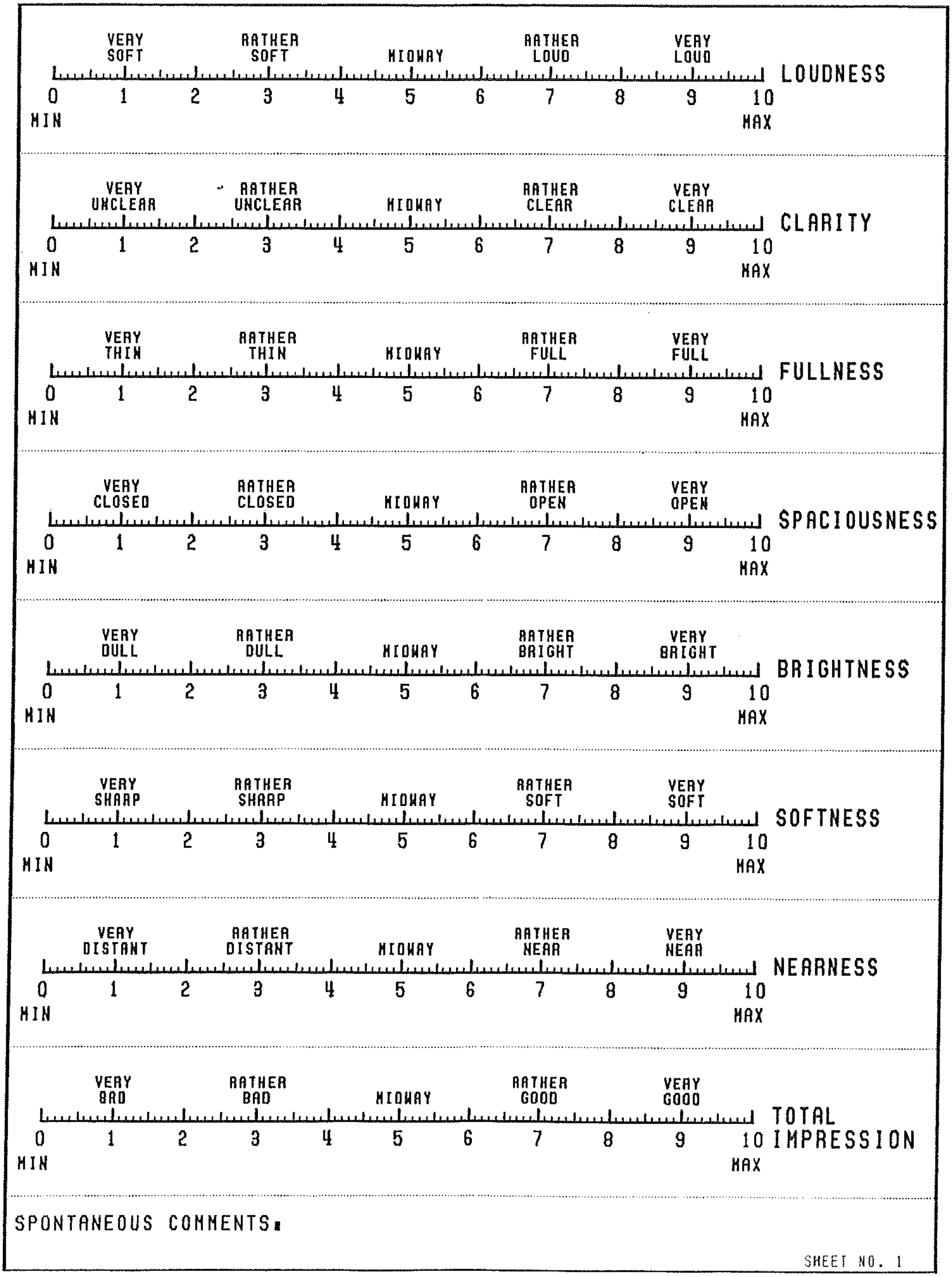

Figure 3b. Example of the response form (translated into English).

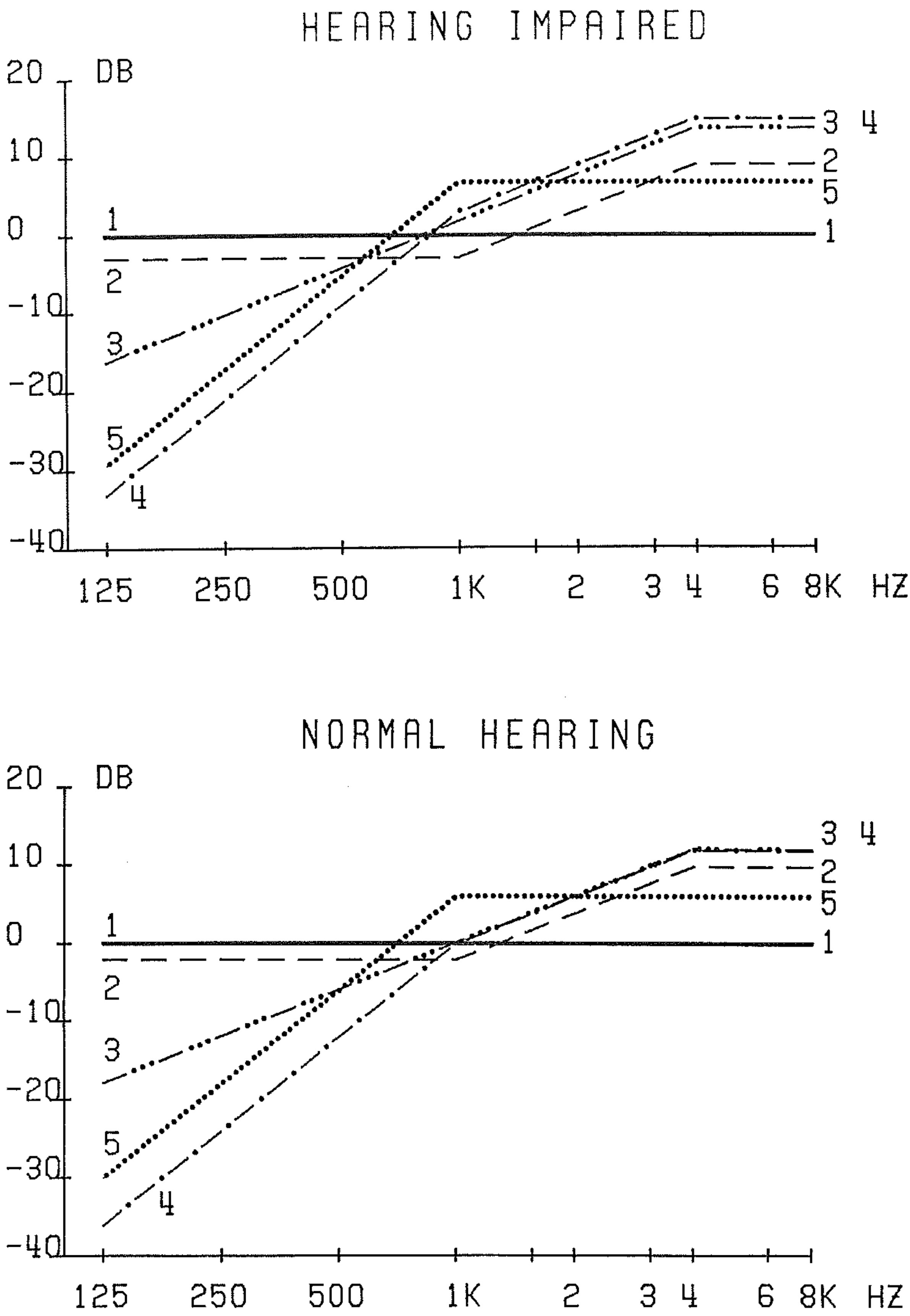

Figure 4. Schematic frequency responses for the five systems including differences in comfortable listening levels used in the experiment.

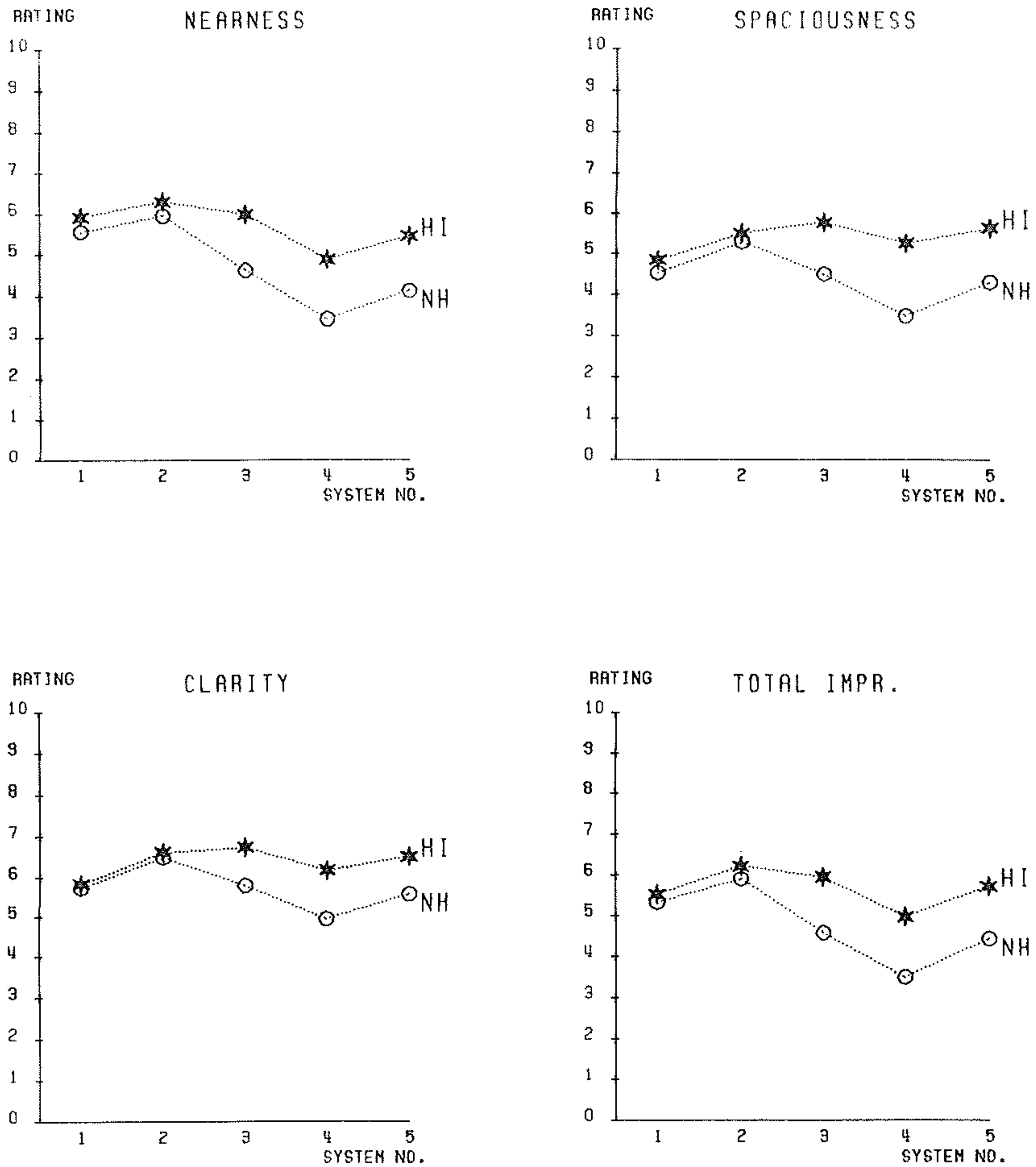

Figure 5. Continued.

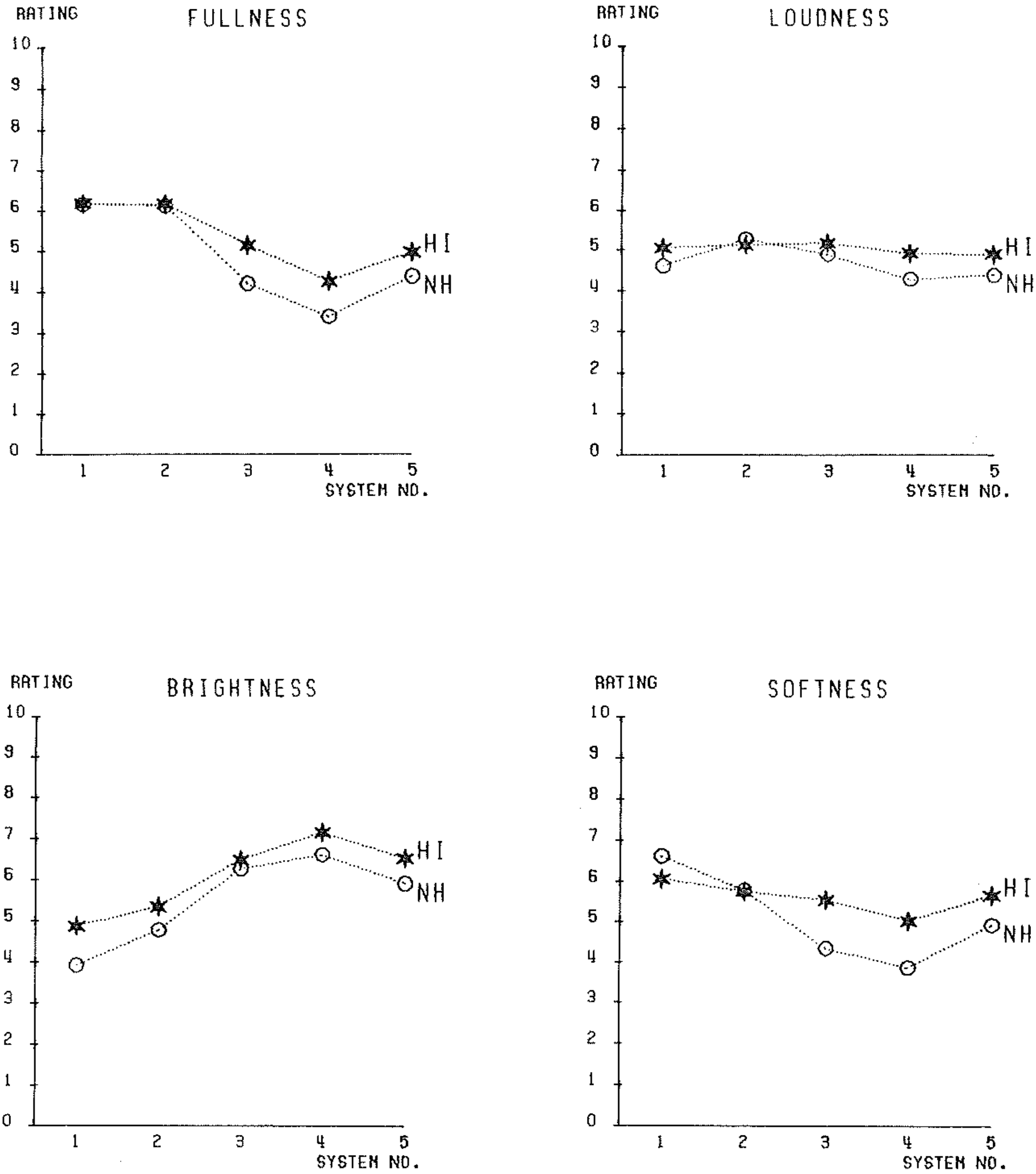

Figure 5. Mean ratings for systems Nos. 1-5 over all programs for the HI and NH groups. (The dotted lines are drawn solely to facilitate the comparison between the two groups.)

Appendix 1. Instructions.

Instructions for judgments of sound quality

Your task today is to judge the sound quality of the programs you listened to in the previous session. You shall now try to describe how they sound by means of the scales that you see on the response form. The scales refer to various properties of the sound reproduction. They are graded from 10 (maximum) to 0 (minimum). You decide yourself on the accuracy that you consider necessary. As you can see it is also possible to use decimals. The integers 9, 7, 5, 3, and 1 are defined on the response form. For instance, in the scale for clarity 10 means maximum (highest possible) clarity, 9 means very clear, 7 rather clear, 5 midway, 3 rather unclear, 1 very unclear, and 0 minimum (lowest possible) clarity. The other scales work in similar ways.

The scales may be further defined as follows:

Clarity: The reproduction sounds clear, distinct, and pure. The opposite is that the sound is diffuse, blurred, thick, and the like.

Fullness: The reproduction sounds full in opposite to thin.

Brightness: The reproduction sounds bright in opposite to dull and dark.

Softness: The reproduction sounds soft and gentle in opposite to sharp, hard, keen, and shrill.

Spaciousness: The reproduction sounds open and spacious in opposite to closed and shut up.

Nearness: The sound seems to be close to you in opposite to at a distance.

Loudness: The sound is loud in opposite to soft (faint).

Total impression: An overall judgment of how good you think the reproduction is.

There is a new response form for each case. First we are going to practice with some programs. Do your rating in each scale without looking at the other scales.

(Compare the corresponding instructions for judgments of loudspeaker reproductions by high fidelity experienced listeners in Gabrielsson & Lindström, 1985).

Instructions for ratings of ideal values

Now you should use the response form to designate how an ideal reproduction of each program should be. You are going to listen once more to each program. But now you should not judge the reproduction you listen to, but instead try to imagine how each program should sound with regard to fullness, loudness, brightness, softness, nearness, spaciousness, and clarity, in order to get an ideal reproduction. (The ideal value in total impression is of course 10, so you can skip this scale.)

You thus designate the ideal value in each of the seven scales above by imagining how an ideal reproduction of each program should sound.

Instructions for listening to sentences in noise

We want to investigate how difficult it is to perceive speech in a background of noise. You are going to hear sentences of five words, for instance, "Bertil got eight white socks". Depending on the background noise it is sometimes easier and sometimes more difficult to perceive the words. Please, repeat clearly the words you have heard. If you want, you may guess, but don't hesitate too long, otherwise you will miss the beginning of the next sentence.