ISSN 0280-6819

TECHNICAL AUDIOLOGY

Report TA No. 117
December 1988

PERCEIVED SOUND QUALITY OF REPRODUCTIONS WITH DIFFERENT FREQUENCY RESPONSES AND SOUND LEVELS

Alf Gabrielsson, Björn Hagerman, Tommy Bech-Kristensen and Göran Lundberg

KAROLINSKA INSTITUTET
TEKNISK AUDIOLOGI

Report TA117
December 1988

PERCEIVED SOUND QUALITY OF REPRODUCTIONS WITH DIFFERENT FREQUENCY RESPONSES AND SOUND LEVELS

Alf Gabrielsson, Björn Hagerman, Tommy Bech-Kristensen and Göran Lundberg

Postadress:
TEKNISK AUDIOLOGI
KTH
S-100 44 STOCKHOLM

Besöksadress:
Visiting address:
Lindstedts väg 1 o. 7
STOCKHOLM

Telefon:
08-11 66 60 (sekr.) 08-791 75 15
08-791 75 16
08-791 75 17
08-791 75 18
08-20 89 23 08-791 75 19

KAROLINSKA INSTITUTET
TEKNISK AUDIOLOGI

Report TA117
December 1988

PERCEIVED SOUND QUALITY OF REPRODUCTIONS WITH DIFFERENT FREQUENCY RESPONSES AND SOUND LEVELS

Alf Gabrielsson, Björn Hagerman, Tommy Bech-Kristensen and Göran Lundberg

ABSTRACT

Three programs (female voice, jazz music, pink noise) were reproduced using four different frequency responses and two different sound levels. Fourteen normal hearing subjects listened to the reproductions in earphones and judged the sound quality on seven perceptual scales (loudness, clarity, fullness, spaciousness, brightness, softness/gentleness, nearness), and a fidelity scale. Significant differences among the reproductions appeared in all scales. Interactions between the reproductions and the programs could be explained by the relations between the spectrum of the programs and the used frequency responses. The results for the noise program were similar to those for the other programs.

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

ISSN 0280-6819

Postadress:
TEKNISK AUDIOLOGI
KTH
S-100 44 STOCKHOLM

Besöksadress:
Visiting address:
Lindstedts väg 1 o. 7
STOCKHOLM

Telefon:
08-11 66 60 (sekr.)
08-791 75 15
08-791 75 16
08-791 75 17
08-791 75 18
08-20 89 23
08-791 75 19

CONTENTS

INTRODUCTION

The perceived sound quality of sound-reproducing systems - such as loudspeakers, headphones, and hearing aids - is multidimensional, that is, it is composed by a number of perceptual dimensions. By means of multivariate methods used in experimental psychology we were able to show that the perceived sound quality can be described in terms of dimensions such as clarity, fullness, brightness versus dullness, sharpness versus softness/gentleness, loudness, spaciousness, nearness, and absence of extraneous sounds (Gabrielsson, 1979a; Gabrielsson and Sjögren, 1979a). Rating scales for these dimensions have been successfully used to provide perceptual descriptions of loudspeakers and other sound-reproducing systems (Gabrielsson and Lindström, 1985; Gabrielsson, 1987; Gabrielsson, Schenkman and Hagerman, 1988). Overall evaluations of the systems in terms of fidelity or pleasantness may be regarded as weighted combinations of the separate perceptual dimensions. The weight given to each dimension - that is, how important it is for the overall evaluation - depends on the character of the program to be reproduced, the listener's earlier experiences, and so forth.

The relations between the perceptual dimensions and various physical properties of the systems are complex and still largely unknown. Among the physical variables the frequency response is often considered as the most important. Its effects on the perceptual dimensions have been explored in a post hoc manner by studying the frequency response of the systems receiving different ratings in the respective dimensions and also more directly by experimental manipulation of the frequency response (Gabrielsson, Rosenberg and Sjögren, 1974; Gabrielsson and Sjögren, 1979a, 1979b; Gabrielsson, Lindström and Till, 1986, 1987; Gabrielsson, Schenkman and Hagerman, 1988; cf. also similar approaches by Komamura, Tsuruta and Yoshida, 1977; Kötter, 1968; Staffeldt, 1974; Toole, 1986). The results from our investigations indicate that the frequency response can affect any of the above-mentioned perceptual dimensions. Brightness and sharpness increase (dullness and softness/gentleness decrease) with rising frequency response toward higher frequencies and/or falling response toward lower frequencies (cf. also Stevens and Davis, 1938 pp. 163-166, for density and brightness in sinusoidal tones; Bismarck, 1974). Fullness is favored by a broad frequency range and relatively more emphasis on lower frequencies (cf. Stevens and Davis, 1938 p. 161, for volume in sinus-

oidal tones). Clarity, spaciousness, and (to some extent) nearness are likewise favored by a broad frequency range, often with a certain emphasis on midhigh to high frequencies. The effects of extraneous sounds, e.g. hiss, may be relieved by reduced response at high frequencies. Generally the results also depend on the characteristics of the (music or speech) programs that are reproduced. There are thus interactions between the reproduction systems and the programs.

Another important physical factor is obviously the sound level. The available evidence (e.g., Gabrielsson and Sjögren, 1979a) indicates that an increase of the sound level will usually increase the perceived fullness, spaciousness, and nearness as well as sharpness and brightness; a decrease of the sound level gives the opposite results. Increased sound level may also contribute to increased clarity, although only up to a certain level at which overloading may occur. There may be interactions between the sound level on the one hand and the frequency response and/or the spectrum of the program on the other hand. For instance, a program reproduced by a system with boosted treble may sound even sharper and brighter if the sound level is increased, while a reproduction with boosted bass will probably sound even duller if the sound level is raised.

Because of such complex interactions and also because of the post hoc character of certain results referred to above a further experiment with systematic manipulation of the frequency response and the sound level was conducted. The purpose was to investigate the effects on the perceptual dimensions of four markedly different frequency responses (flat, boosted at low, midhigh, and high frequencies) at two different sound levels in the reproduction of three programs including speech, music, and pink noise. Hypotheses concerning the effects were stated on the basis of results from our earlier investigations and are given below under Results.

METHODS

Programs

Three programs were used:

1. Pink noise (-3dB/octave), monophonic recording.

2. Female voice reading a fairy-tale in an anechoic chamber, monophonic recording.

3. Jazz music, excerpt from "Ole Miss" by W.C. Handy, performed by The Peoria Jazz Band in an auditorium. Phonograph record: Opus 3, 79-00, Testskiva 1: Perspektiv. The excerpt was copied directly from the stereophonic recording on the master tape but was played monophonically to the listener.

Each program lasted for about one minute.

The pink noise was chosen to serve as a neutral reference. The female voice in the anechoic chamber has most of its energy below 1 kHz, especially between about 130 and 700 Hz, while the jazz music has a considerably broader frequency range with a boost around 100 Hz, see Figure 1., page 17. The programs were low-passed at 6.7 kHz for reasons explained below.

Reproduction system

The reproduction system is displayed in Figure 2., page 18. A tape recorder Telefunken Magnetophon 28 was used to reproduce the programs, which were then filtered before binaural presentation to the listeners through Sony Walkman MDR-E262 earphones. The frequency response of the earphones is shown in Figure 3., page 19. The steep cut-off at about 6 kHz is due to the anti-alisasing low-pass filter described below.

The filters were implemented as digital FIR-filters using a TAMP3 equipment. TAMP3 - Technical Audiological Measuring Processor, revision 3 - is a general purpose measuring device developed in our department. It can be used for measuring the frequency response of linear, time-invariant systems and also for digital filtering of signals in real time. The main part of TAMP3 is a processor chip, TMS32010, specially made for digital signal processing purposes. TAMP3 is equipped with anti-aliasing filters and fast AD and DA converters. Various types of input amplifiers, output

amplifiers and attenuators can be connected to the processor. The equipment is contained in an industry standard 19" rack and controlled by an ABC 808 microcomputer.

The sampling frequency of the digital filter had to be restricted to 20 kHz, and an anti-aliasing lowpass filter was set to 6.7 kHz. Four different filters were implemented. One filter had a flat response, that is, no filtering at all. The other three filters meant about 10 dB amplification below 200 Hz, around 1 kHz, and around 4 kHz, see Figure 4., page 20. (The lowest filter could not be made symmetrical because of certain limitations in the equipment. Below 100 Hz there is anyhow little energy due to the cutoff of the earphones, cf. Figure 3.) In the following these filters will be referred to as the L (for low), M (midhigh), and H (high) filter, respectively.

The sound levels were set to represent an approximately natural level of the respective program when listened to in the earphones with no filtering. Measured by a coupler according to IEC 711 fitted into the KEMAR manikin, the A-weighted sound level for the pink noise with the flat response was about 68 dB, for the female voice about 56 dB, and for the jazz music about 80 dB. For comparison each program was also presented at a 10 dB lower level.

The filters themselves caused certain changes in the sound level. These effects were different for different programs depending on their spectrum (Figure 1). For the pink noise there was practically no difference in the A-weighted sound level between the flat response and the L filter, while the M filter meant an increase by 2 dB, and the H filter an increase by 6 dB. For the jazz music there was again no difference in the A-weighted sound level between the flat response and the L filter, while the M filter meant an increase of about 5 dB and the H filter an increase of about 3 dB. For the female voice the L filter increased the A-weighted sound level by about 3 dB, the M filter about 2 dB, and the H filter about 1 dB. These effects have to be considered in the interpretation of the results.

The listener was seated in a sound insulated chamber used for psychoacoustic experiments. All equipment and the experimenter were in an adjoining room.

Subjects

Fourteen subjects, 7 males and 7 females, age 22-34 years participated. None of them had any experience from this type of experiment. All of them were tested for normal hearing (less than 20 dB hearing loss 250 8000 Hz, ISO 389).

Response variables

The reproductions were rated in eight scales. Seven of them refer to perceptual dimensions: loudness (Swedish: ljudstyrka), fullness (fyllighet), brightness (ljushet), softness/gentleness (mjukhet), nearness (närhet), spaciousness (rymdkänsla), and clarity (tydlighet). The eighth scale meant an overall evaluation of each reproduction in terms of its fidelity. All scales were graded from 10 (maximum) to 0 (minimum) and with definitions for 9, 7, 5, 3, and 1 as seen in Figure 5., page 21. Decimals were included, since many subjects in earlier investigations used decimals in their ratings (Gabrielsson and Lindström, 1985; Gabrielsson et al., 1988). Further explanations were given in the instructions, see Appendix.

Design and procedure

There were in all 24 stimuli, 3 programs x 4 filters x 2 sound levels, and they were rated twice by each subject in all eight scales (however, the noise program was rated in seven scales omitting the fidelity scale). The presentation order of the stimuli was randomized, differently for each subject. The order of the rating scales on the response form (Figure 5) was also randomized differently for each subject. After introducing the subject to the situation and trying out the earphones, the instructions (Appendix) were given followed by 12 practice trials. Then the main experiment including 48 trials (24 stimuli x 2 trials per each stimulus) was conducted with a break in the middle. After that the subject answered some questions related to the experiment.

In order to check that the earphones were placed in an identical manner in the two parts of the main experiment (before and after the break), a broadband signal was fed to the ear, and the resulting response was measured by a Diaphon probe microphone inserted into the ear canal behind the earphone. The microphone response was fed into TAMP3 for analysis of the frequency response. On the whole these frequency responses were

fairly similar in both parts of the experiment. Between 200 and 3000 Hz the differences were usually less than 2 dB. A dip was often found around 150 Hz, somewhat varying in position and size. Differences larger than 5 dB were found around 4-5 kHz in a few cases. Since the variance of the repeated ratings for each stimulus was comparable to that found for a group of corresponding subjects in an earlier experiment (Gabrielsson et al.,1988; cf. under Results below), there is no reason to believe that there were larger differences in the placement of the earphones before and after the break than what can be considered as natural and inevitable.

The total time required for each subject was about 2.5 hours. Beside the necessary time for the instructions, the practice trials, and the break, much time was spent in fitting the probe microphone and the earphone. The actual listening time was about 50 minutes.

Data treatment

The subjects' ratings were subjected to analysis of variance, separately for each scale. This was done both for each individual subject (sources of variance: filters, sound levels, and programs; fixed model) and over all subjects (sources: filters, sound levels, programs, and subjects; mixed model). One-tailed *t* tests were used to test specific hypotheses concerning the effects of different filters. For general principles concerning analysis of variance and related questions see Winer (1971) or Kirk (1982), and for application in listening tests Gabrielsson (1979b).

RESULTS

Reliability of ratings

The intra-individual reliability was studied by means of the "within cell mean square" (MSw) in the individual analyses of variance, that is, the estimated average variance of the two ratings made for each stimulus in each scale (MSw is the error term for the *F* tests in the fixed model). The smaller this variance, the better is of course the reliability. The median value for MSw over all 14 subjects for each scale appears in Table 1. (The median was chosen rather than

the arithmetic mean because of one extremely deviating subject.) MSw is clearly lowest for the loudness scale (0.53), which is the most familiar dimension. For the other scales MSw varies between 1.27 and 1.64. These values are about the same as for another group of unselected subjects with normal hearing in Gabrielsson et al.(1988), but higher (that is, the reliability is worse) than for subjects selected for experience of listening to high fidelity sound reproduction (Gabrielsson and Lindström, 1985). Another indication of good reliability is the occurrence of significant F tests (at .05 level or lower) for the different experimental variables. Out of the 14 subjects typically at least half of them had significant differences among the various filtered reproductions in each scale.

The inter-individual reliability (the agreement between the subjects) was estimated by means of the $\underline{r}_b$ index (Winer, 1971, p. 283; Gabrielsson, 1979b). Its maximum value is 1.00; the higher, the better reliability. As seen in Table 1, the reliability is again highest for loudness (0.98), but is generally high (0.84-0.95) for the other scales as well.

Table 1.

Median value across subjects for MSw and value of the $\underline{r}_b$ index for each rating scale.

	MSw	$\underline{r}_b$
Loudness	0.53	0.98
Clarity	1.40	0.91
Fullness	1.37	0.86
Spaciousness	1.64	0.91
Brightness	1.27	0.92
Softness	1.27	0.95
Nearness	1.48	0.93
Fidelity	1.29	0.84

Effects of filters and sound levels

Each program was reproduced in eight different ways, using 4 filters x 2 sound levels. The average ratings across all subjects in each scale for the different reproductions are shown in Figure 6., page 22-26. The

hypotheses and the results will be discussed separately for each scale.

Loudness

It was of course expected that the loudness ratings would reflect the two different sound levels as well as the level settings of the programs. There was a highly significant difference between the two sound levels, $F(1, 13) = 140$, $p < .001$. As seen in Figure 6, the ratings at the high sound level are 1.5 - 3 units higher than for the corresponding cases at the low level. There was also a significant difference in rated loudness among the programs, $F(2, 26) = 9.9$, $p < .001$, accompanied also by a significant program x level interaction, $F(2,26) = 18$, $p < .001$. The meaning of these results is clear from Figure 6. At the high level the voice is rated lower in loudness than the other programs, while there is practically no difference among the programs at the 10 dB lower level. The perceived loudness is thus reduced more for the noise and jazz programs than for the voice, when the sound level is lowered. The fact that the noise is rated almost as high in loudness as the jazz music at the high level, although there is a considerable difference between their sound levels, is probably due to the continuous and "irritating" character of the noise.

There was also a significant difference among the various filters, $F(3, 39) = 20$, $p < .001$, accompanied by a significant filter x level interaction, $F(3, 39) = 6.2$, $p < .01$. As seen in Figure 6 for the high level, all three filters seem to increase loudness in comparison with loudness for the flat response, which may be expected since the filters in most cases also introduce a certain increase of the sound level. An exception is the H filter at the voice program; but since the voice has most of its energy below 1000 Hz, it is not much affected by the H filter (cf. Figures 1 and 4). At the low level there are similar but less pronounced tendencies.

Clarity

It was expected that the high, "natural" sound level would provide better clarity than the lower level. This was confirmed by the corresponding F test, $F(1, 13) = 19.8$, $p < .001$, and is evident in Figure 6. The difference is mostly 1.0 - 1.5 units. There was also a significant difference among the programs, $F(2, 26) = 9.3$, $p < .001$, meaning that the noise is

rated lower in clarity (which seems natural) than the other two programs, among which the jazz music is usually rated higher than the voice.

With regard to the filters it was expected that the L filter would reduce clarity due to more masking by low frequency components, while the M and H filters would increase clarity in comparison with the flat response. As seen in Figure 6 the results mainly agree with these hypotheses. The average rating for the L filter across all programs and both sound levels (5.0) is lower than the corresponding value for the flat response (5.6), $t(39) = 2.4$, $p<.025$. Likewise the average rating for the H filter (6.2) is higher than for the flat response, $t(39) = 2.4$, $p<.025$. The difference between the M filter and the flat response does not reach statistical significance. However, the difference between the M filter and the L filter is significant, $t(39) = 3.2$, $p<.005$. These results look similar at both sound levels and for all programs, including the "neutral" noise program, see Figure 6.

Fullness

It was hypothesized that the higher sound level would provide more fullness than the lower level, and that the L filter would increase perceived fullness in comparison with the flat response. Both hypotheses were confirmed as can be seen in Figure 6. The difference between the two sound levels is significant, $F(1,13) = 17.0$, $p < .01$. The difference between the ratings for the L filter and the flat response across programs and sound levels is also significant, $t(39) = 2.65$, $p < .01$.

There is a tendency, just short of statistical significance, that the H filter gives less fullness than the flat response at the higher sound level, especially for the noise and jazz programs. The reason is probably that the introduction of the H filter makes the lower frequencies, which contribute most to fullness, less important. For the voice program the effect of the H filter is small, since most of its energy lies below the range of the H filter.

Spaciousness

The higher sound level was expected to provide more perceived spaciousness than the lower level. This is also the case as seen in Figure 6 and confirmed by the statistical test for the difference between the two

sound levels, $F(1,13) = 35.2$, $p < .001$. It is further evident from the figure that the jazz program sounds more spacious than the other two programs, especially in comparison with the voice program. This is an effect of the respective recordings: in a large auditorium for the jazz program but in an anechoic chamber for the voice program.

With regard to the effects of the filters it was expected that spaciousness would possibly increase with the M and H filters and/or decrease with the L filter in comparison with the flat response. The former part of the hypothesis was not confirmed, while the latter was. The difference between the L filter and the flat response across programs and sound levels was significant, $t(39) = 1.72$, $p < .05$. An exception occurs for the jazz program at the lower sound level, where there is no difference between the flat response and the L filter.

Brightness

According to our hypothesis perceived brightness should decrease with the L filter but increase with the M and especially with the H filter in comparison with the flat response. As seen in Figure 6, the L filter reduced the brightness throughout, and the difference between the flat response and the L filter across programs and sound levels is strongly significant, $t(39) = 6.04$, $p < .0005$. The H filter increased brightness, and the difference between the H filter and the flat response is also highly significant, $t(39) = 3.04$, $p < .005$. There is no significant difference between the M filter and the flat response, although there is a tendency in the expected direction at the higher sound level.

It can be noted that the difference between the H filter and the flat response is much smaller for the voice program than for the other two programs. The reason is that the H filter does not influence the voice program very much, since it lies mainly outside the spectrum of the voice program.

It was also tentatively hypothesized that brightness would be higher for the higher sound level than for the lower level. As seen in Figure 6, this expectation was not confirmed.

Softness/Gentleness

It was expected that the lower sound level would sound softer/more gentle than the higher sound level, or, in other words, that the higher sound level would sound sharper than the lower. This was confirmed as seen in Figure 6 and by the statistical test of the difference between the two sound levels, $F(1,13) = 47.2$, $p < .001$. There are also significant differences in softness among the programs, $F(2,26) = 7.6$, $p < .01$, accompanied by a significant program x level interaction, $F(2,26) = 12$, $p < .001$. As seen in Figure 6, the voice program sounds softer than the noise and the jazz music, and the differences among the programs are more pronounced at the higher level than at the lower.

With regard to the filters it was expected that the L filter would increase softness, while the M and H filters would decrease softness in comparison with the flat response. The data in Figure 6 indicate more softness for the L filter than for the flat response, but the difference does not reach conventional statistical significance, $t(39) = 1.08$, $p < .15$. The difference between the M filter and the flat response is significant in the expected direction, $t(39) = 2.74$, $p < .005$, and this is also true for the difference between the H filter and the flat response, $t(39) = 3.33$, $p < .005$. There was also a significant level x filter interaction, $F(3,39) = 5.2$, $p < .01$, meaning that the decrease of softness with the M and H filters is more pronounced at the higher level than at the lower, see Figure 6. In our own experience the noise and the jazz music sound sharp and irritating, when reproduced by the H filter at the high level.

There is finally a significant filter x program interaction, $F(6,78) = 6.5$, $p < .001$. While the noise and jazz music sounds sharpest with the H filter, the voice program sounds sharpest with the M filter, see Figure 6. The H filter cannot contribute very much to sharpness in the voice program, since it lies higher in frequency than the main part of the voice spectrum.

Nearness

The main hypothesis was that the higher sound level would give more impression of nearness than the lower level. This was confirmed as seen in Figure 6 and by the statistical test of the difference between the two sound levels, $F(1,13) = 95$, $p < .001$. There was also a significant difference among the programs, $F(2,26) = 5.7$, $p < .01$, meaning that the voice program

sounds nearest (it is recorded in an anechoic chamber) and the "neutral" noise program sounds most distant.

Regarding the filters it was possibly expected that the M and H filters would contribute to increased nearness in comparison with the flat response. Although there are some tendencies in this direction, the results are not consistent and the corresponding statistical tests are not significant. It may be noted that the noise and jazz programs tend to sound nearer for all three filters, while this does not hold for the voice program.

Fidelity

It was expected that the fidelity should be better for the higher sound level, since this was set by the experimenters to approximately correspond to the original level at the recordings. This hypothesis was confirmed as seen in Figure 6, and the statistical test for the difference between the sound levels was significant, $F(1,13) = 7.6$, $p < .025$.

It was of course also expected that fidelity would be affected by the different filters, especially by the L filter, since an emphasis on low frequencies tends to introduce much masking of higher frequency components. As seen in Figure 6, the fidelity is throughout worst for the L filter, and the difference between the flat response and the L filter is highly significant, $t(39) = 3.84$, $p < .0005$. However, there is also an interaction between sound levels and filters, $F(3,39) = 4.7$, $p < .01$. The differences among the filters are much more evident at the higher, natural sound level than at the lower level. At the lower level the only clear result is that the L filter is worse than the others. However, at the higher level the L, M, and H filters are all worse than the flat response.

At the "natural" level then the flat response is superior to all other reproductions. This may seem to contradict earlier results indicating that a slight or moderate emphasis on midhigh to high frequencies is favorable to the impression of fidelity (Gabrielsson and Sjögren, 1979a; Gabrielsson et al., 1988). However, the present M and H filters represent a very

pronounced boost (up to 10 dB) of certain frequency ranges, which evidently may counteract good fidelity.

DISCUSSION

The manipulations of the sound level and/or the frequency response affected all perceptual dimensions included here. The results mainly agree with what could be expected from our earlier investigations (see Introduction). The higher, natural sound level provided better clarity, more fullness, spaciousness, and nearness - but less softness/gentleness - as well as better fidelity than the 10 dB reduced level. There was no difference with regard to brightness. Use of the L filter resulted in more fullness and softness/gentleness, but less clarity, spaciousness, and brightness (= more dullness), and further worse fidelity in comparison with the flat response. The M and/or H filter resulted in better clarity and more brightness, but less softness/gentleness (= more sharpness) and possibly less fullness in comparison with the flat response. With regard to spaciousness and nearness there were no quite consistent effects of the M and H filters. In fidelity the results were different at the different sound levels. At the lower level the reproductions by M and H filters were rated as about equivalent to that of the flat response, but at the higher, natural level they were rated worse.

These results must be qualified with regard to some interactions between the filters and other factors. There was often an interaction between filters and programs such that the effect of the H filter was different for the voice program than for the other programs, obviously due to the restricted frequency range of the voice program. There were also interactions with the sound levels, meaning that the effects of the filters and/or the difference among the programs were more pronounced at the higher sound level than at the lower; see the results for loudness, fullness, softness, and fidelity in Figure 6.

Interestingly the results for the "neutral" noise program are similar to those for the other programs. The tendencies are in fact very similar for the noise and the jazz music, as seen in Figure 6. Those two programs are also rather similar in their long time average spectrum, see Figure 1.

Since the use of the filters also introduced certain increases of the sound level (see under Reproduction

system in Methods), it is hard to separate the effects of the changed frequency responses from the concomitant changes in the sound level. Furthermore these changes in sound level are different for the different programs depending on their spectrum in relation to the used filter. However, when the 10 dB difference between the two sound levels used here has no effect, such as was the case for brightness, the results for the different filters can be ascribed to the differences in frequency response rather than to differences in sound level. In the remaining (most) cases the situation is ambiguous. Scrutinizing the results regarding the possibility to explain them as due to the increased sound levels associated with the different filters gives no clear answer in either direction. For the present this question is therefore left open.

ACKNOWLEDGMENTS

We wish to express our gratitude to Ove Till and to our subjects. This work was supported by The Bank of Sweden Tercentenary Foundation and the Swedish Ministry of Health and Social Affairs, Commission for Social Research.

REFERENCES

Bismarck, G. von (1974). "Sharpness as an Attribute of the Timbre of Steady Sounds," Acustica **30**, 159-172.

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 (Karolinska Institute, Stockholm).

Gabrielsson, A. (1987). "Planning of Listening Tests: Listener and Experimental Variables," in Perception of Reproduced Sound, edited by S. Bech and O.J Pedersen (Technical University of Denmark, Copenhagen), pp. 51-60.

Gabrielsson, A., and Lindström, B. (1985). "Perceived Sound Quality of High-Fidelity Loudspeakers," Journal of Audio Engineering Society 33, 33-53.

Gabrielsson, A., Lindström, B., and Till, O. (1986). "Loudspeaker Frequency Response and Perceived Sound Quality: Measurements in Listening Room," Technical Audiology Reports No. 114 (Karolinska Institute, Stockholm).

Gabrielsson, A., Lindström, B., and Till, O. (1987). "Loudspeaker Frequency Response and Perceived Sound Quality: Comparison between Measurements in Listening Room, Anechoic Room and Reverberation Room," Technical Audiology Reports No. 115 (Karolinska Institute, Stockholm).

Gabrielsson, A., Rosenberg, U., and Sjögren, H. (1974). "Judgments and Dimension Analyses of Perceived Sound Quality of Sound-Reproducing Systems," Journal of the Acoustical Society of America 55, 854-861.

Gabrielsson, A., Schenkman, B.N., and Hagerman, B. (1988). "The Effects of Different Frequency Responses on Sound Quality Judgments and Speech Intelligibility," Journal of Speech and Hearing Research 31, 166 177.

Gabrielsson, A., and Sjögren, H. (1979a). "Perceived Sound Quality of Sound-Reproducing Systems," Journal of the Acoustical Society of America 65, 1019-1033.

Gabrielsson, A., and Sjögren, H. (1979b). "Perceived Sound Quality of Hearing Aids," Scandinavian Audiology 8, 159-169.

Kirk, R.E. (1982). Experimental Design: Procedures for the Behavioral Sciences (Brooks/Cole, Belmont, California), 2nd ed.

Komamura, M., Tsuruta, K., and Yoshida, M. (1977). "Correlation between Subjective and Objective Data for Loudspeakers," Journal of the Acoustical Society of Japan 33, 103-115.

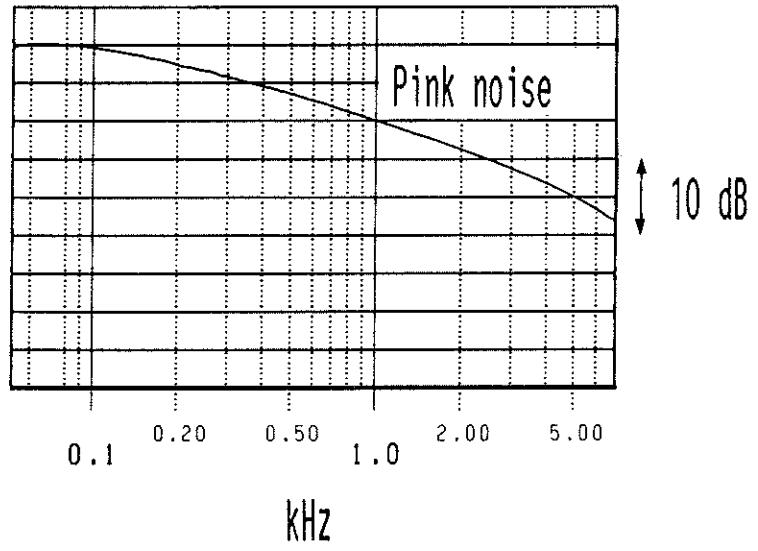
Kötter, E. (1968). Der Einfluss uebertragungstechnischer Faktoren auf das Musikhören (Arno Volk, Köln).

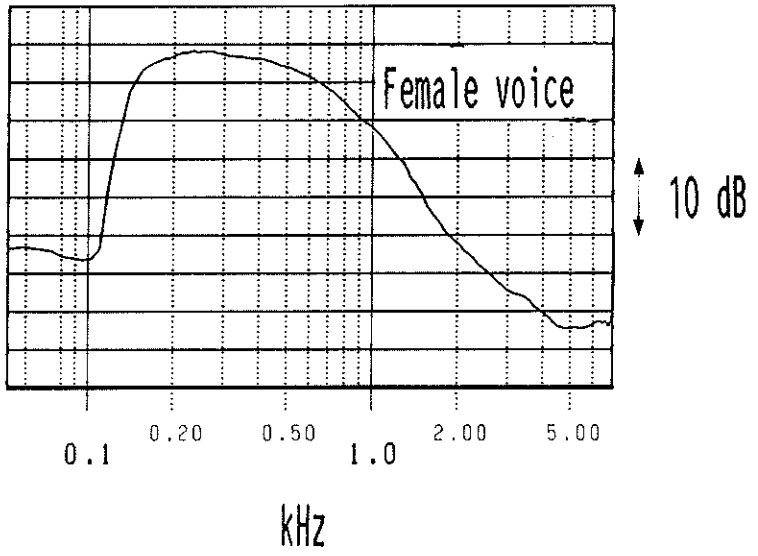
Staffeldt, H. (1974). "Correlation between Subjective and Objective Data for Quality Loudspeakers," Journal of the Audio Engineering Society 22, 402-415.

Stevens, S.S., and Davis, H. (1938). Hearing: Its Psychology and Physiology (Wiley, New York).

Toole, F. (1986) "Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2," Journal of the Audio Engineering Society **34**, 323-348.

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





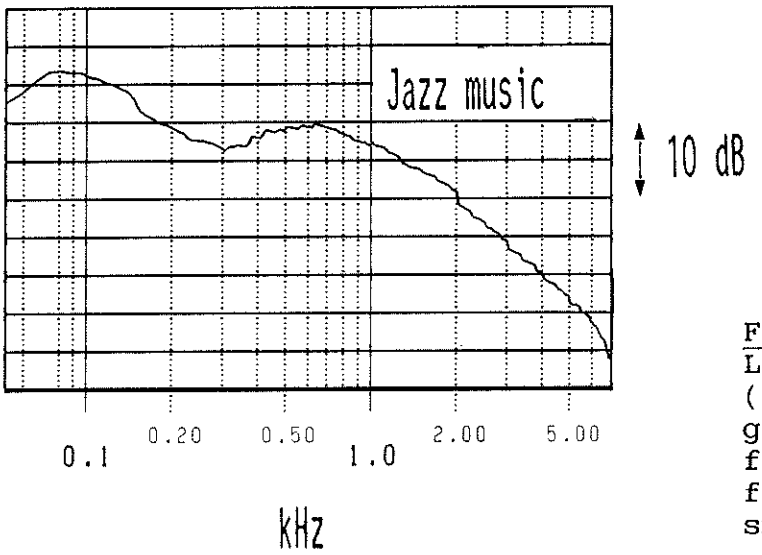


Figure 1.
Long Time Average Spectrum (LTAS) for the three programs. LTAS was calculated from the autocorrelation function of the program and smoothed across octaves.

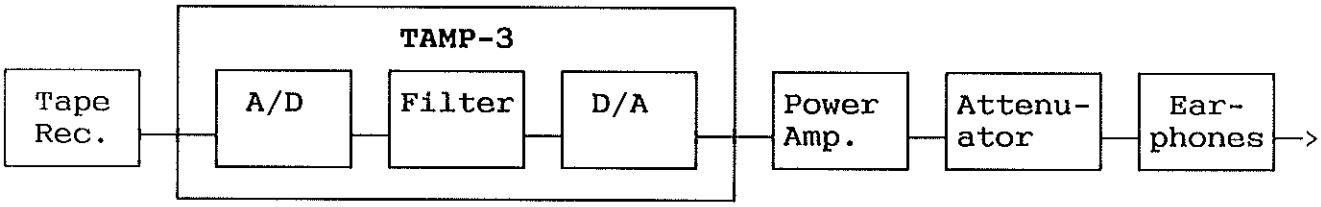


Figure 2. Setup of the reproduction system.

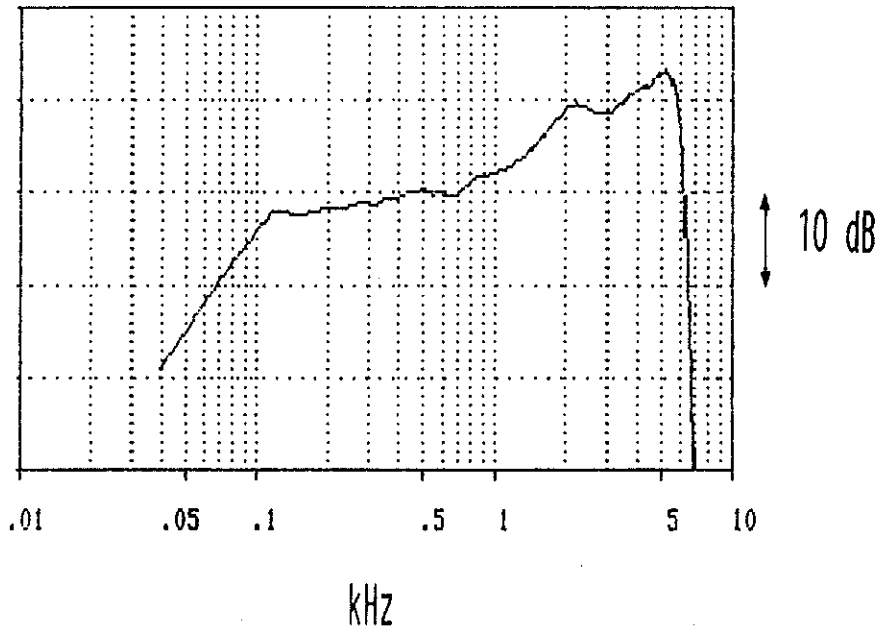


Figure 3. Frequency response of the earphone measured on a manikin (KEMAR) equipped with an ear simulator according to IEC 711.

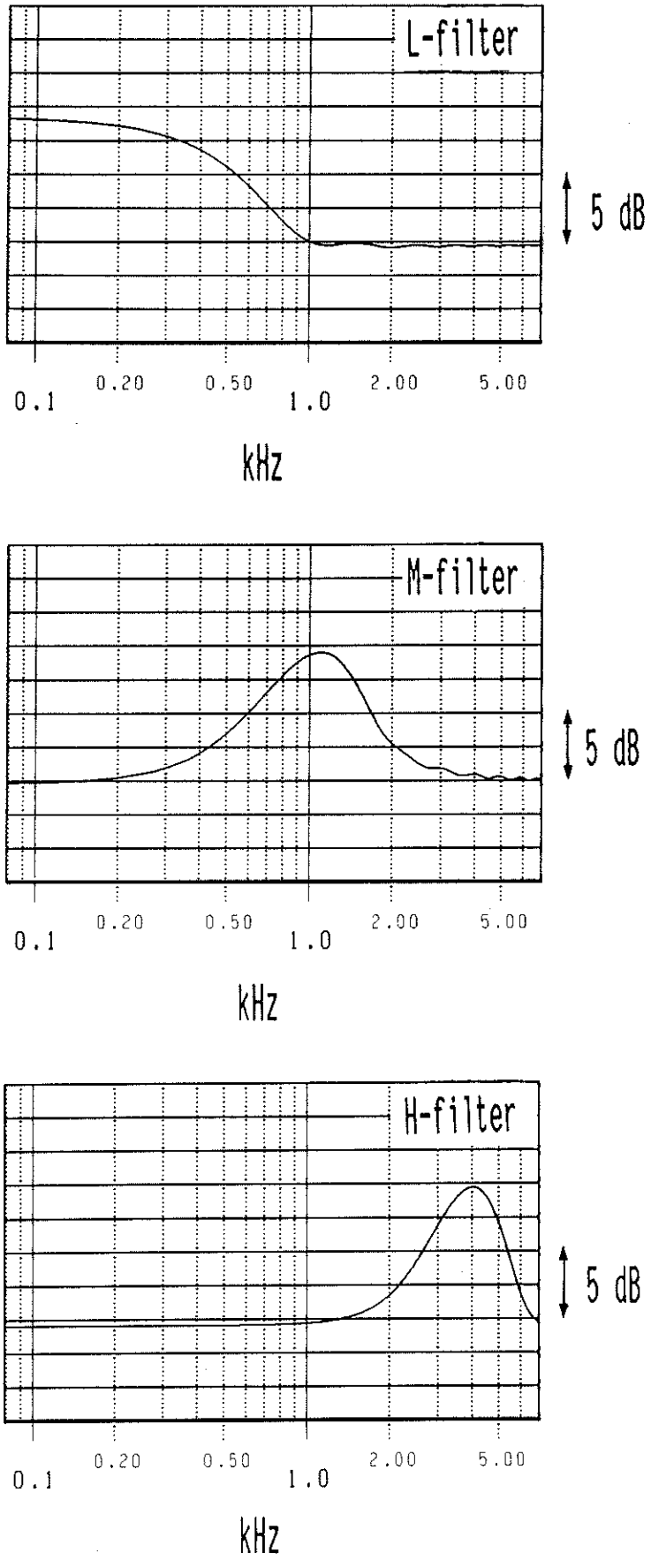


Figure 4.
Frequency responses of the three filters.

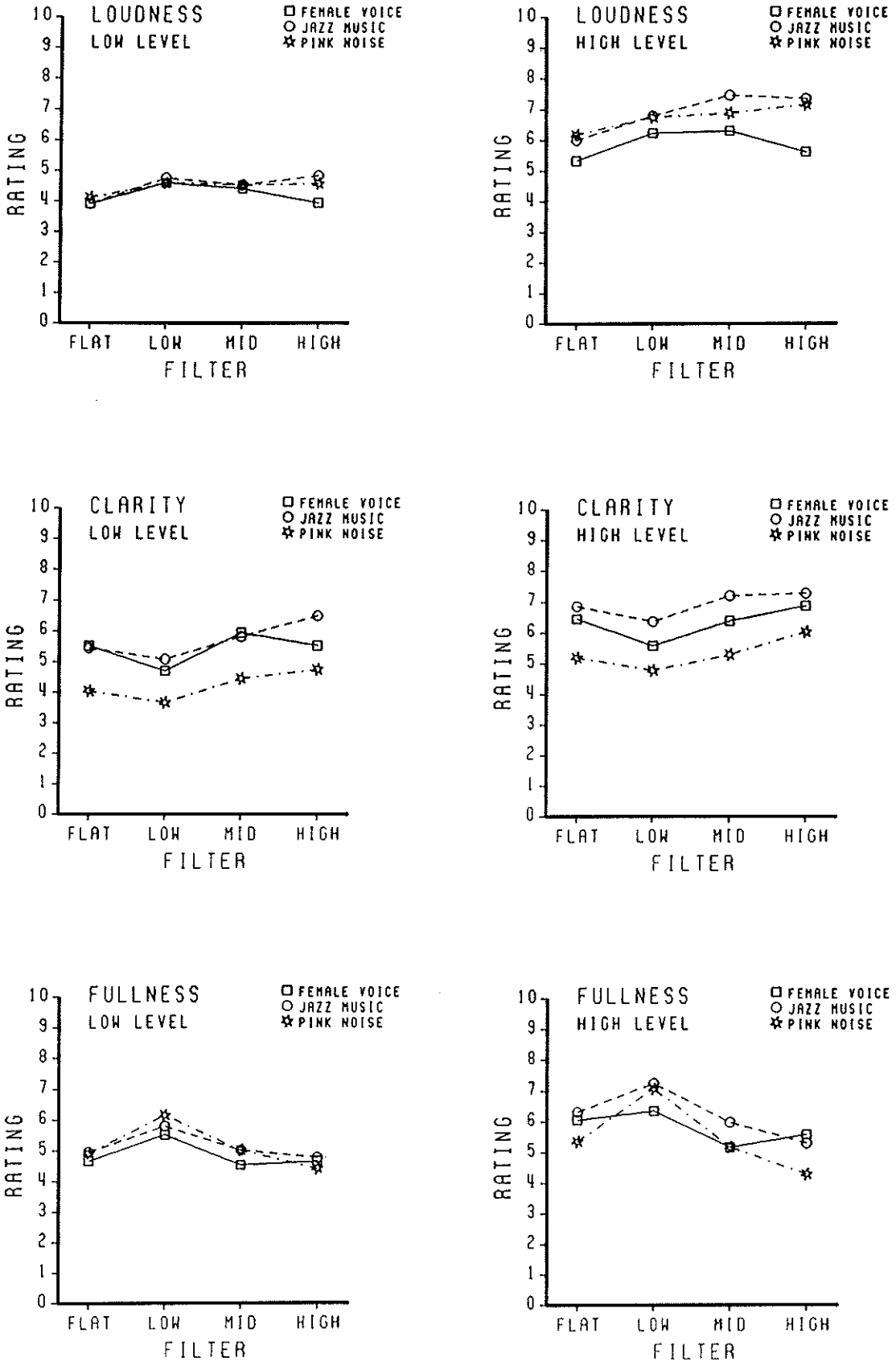


Figure 6. Average ratings across subjects in the different scales for the filtered reproductions at the two sound levels.

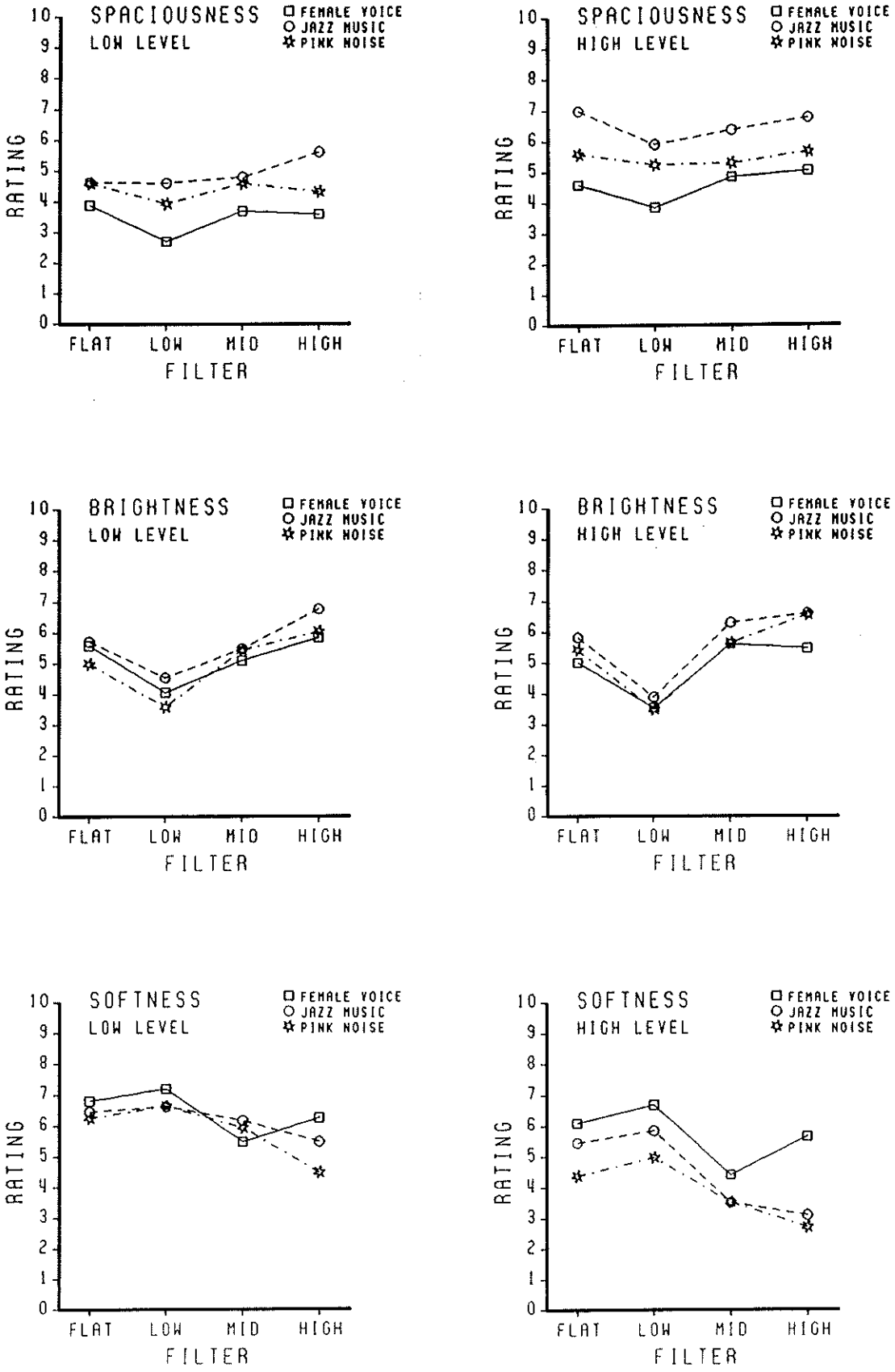


Figure 6. (Continued)

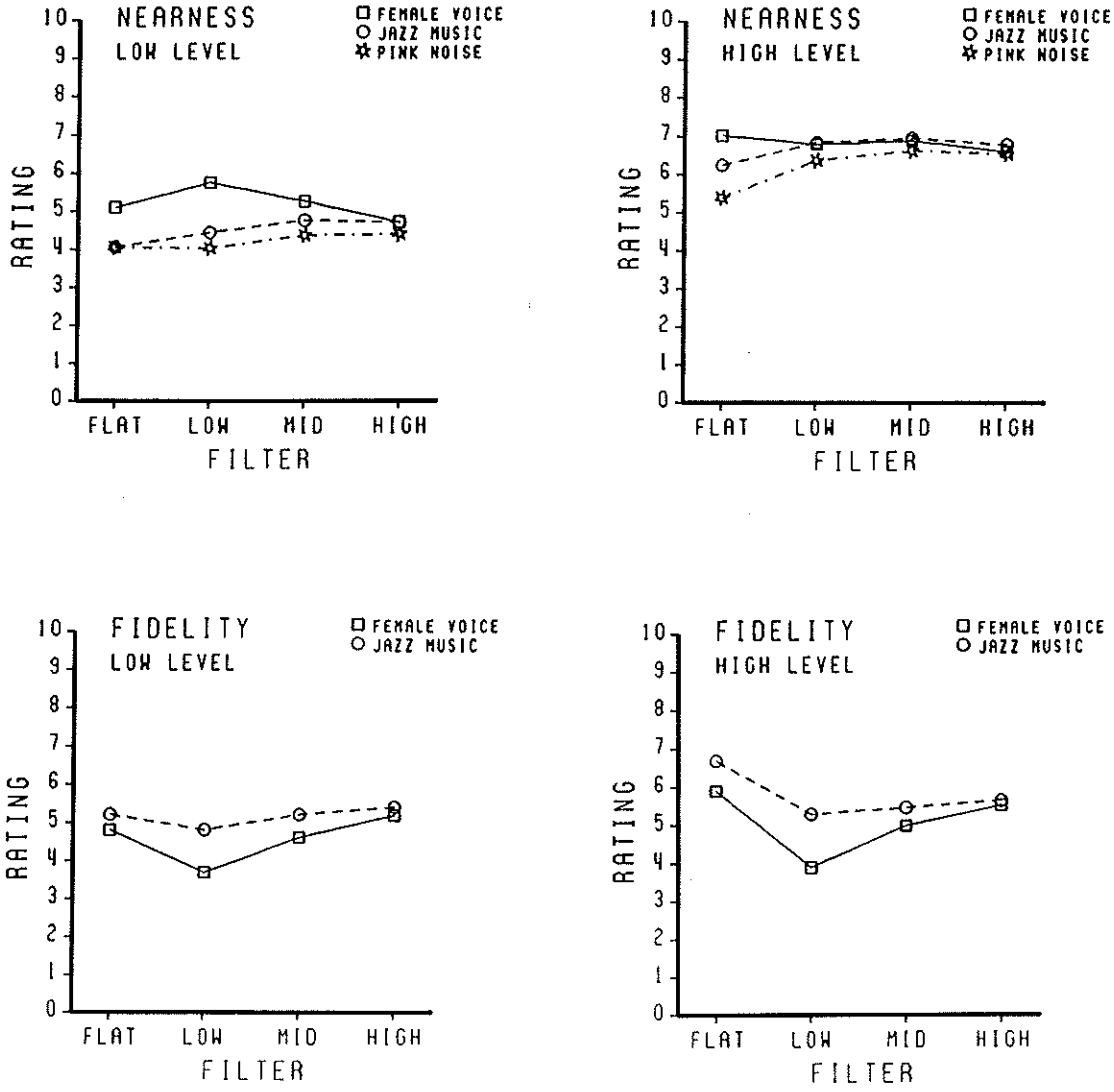


Figure 6. (Continued)

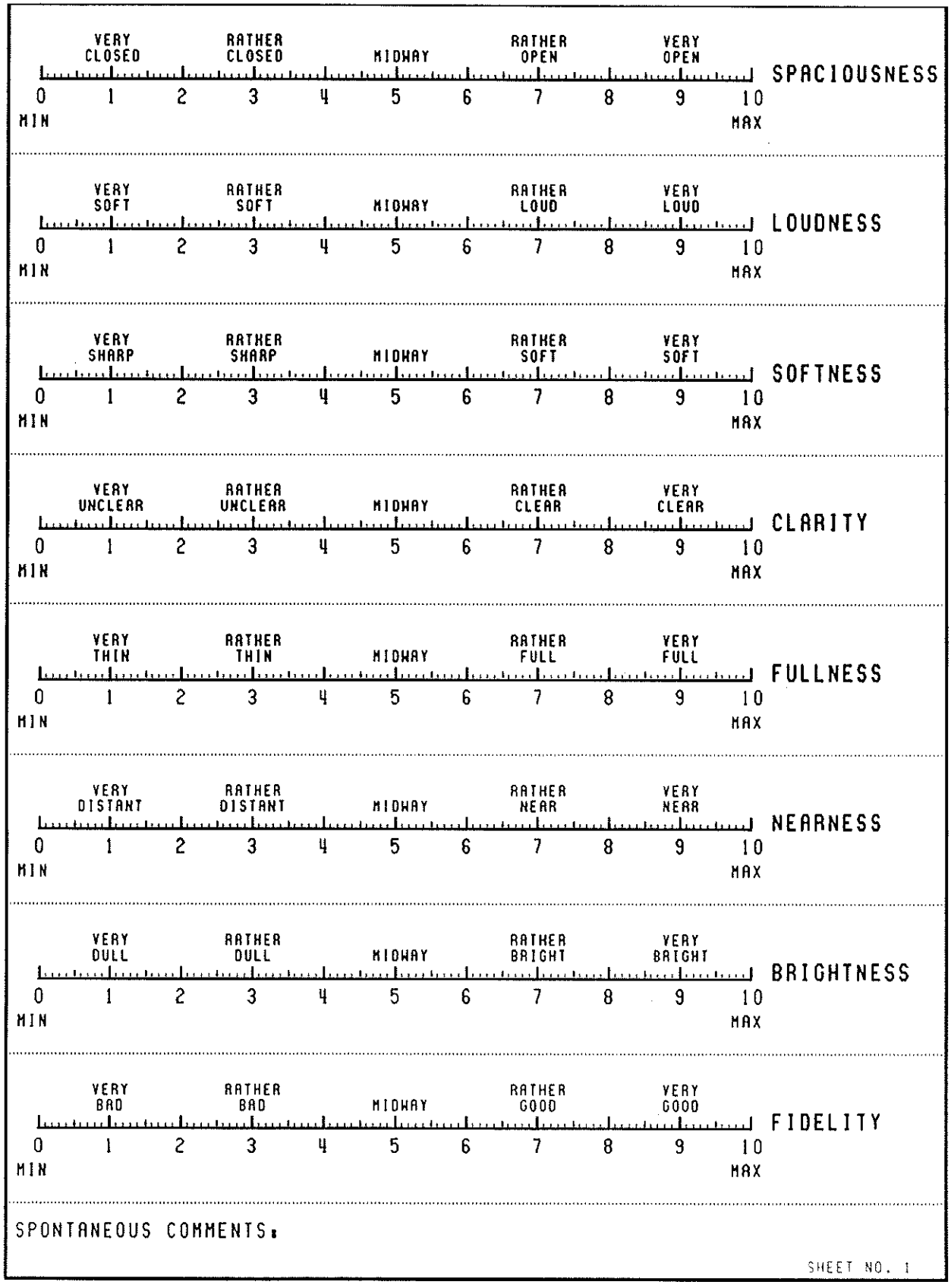


Figure 5. Example of the response form (translated from Swedish).

APPENDIX

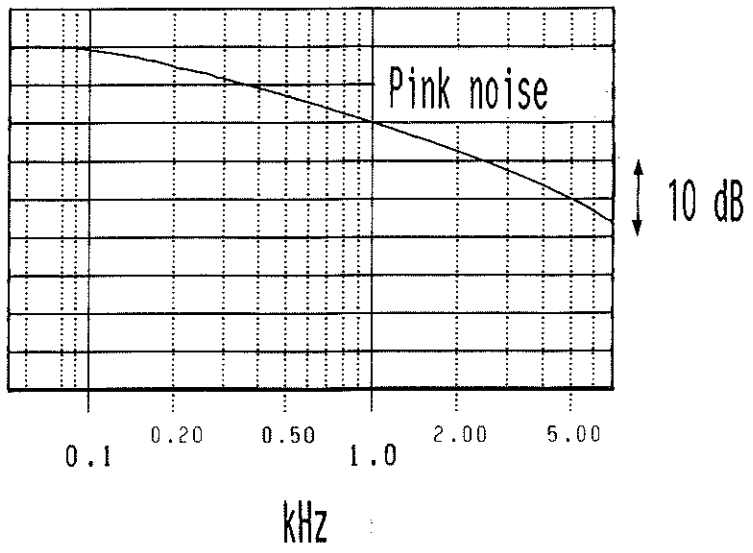
Instructions

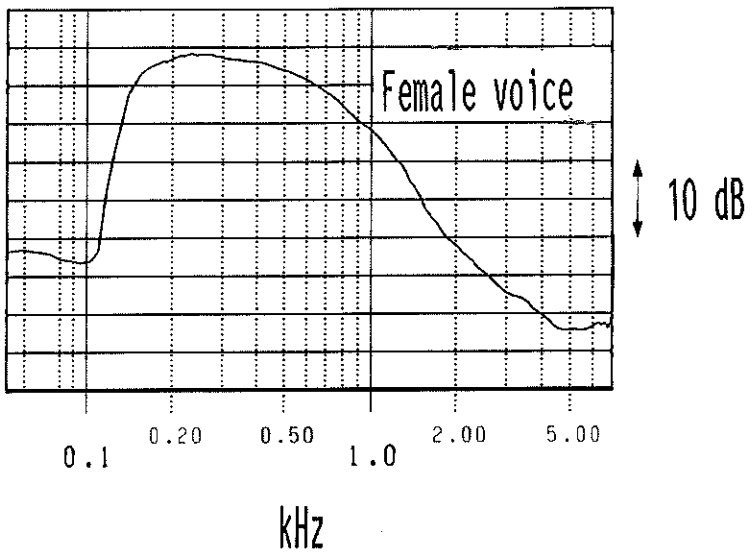
You are going to listen to various reproductions of speech, music, and noise through earphones. Your task is to judge the sound quality of the different reproductions by means of the scales on the response form. The scales refer to various properties of the sound reproduction. They are all graded from 10 (maximum) to 0 (minimum). For instance, in the scale for fullness 10 means maximum (highest possible) fullness, 9 = very full, 7 = rather full, 5 = midway, 3 = rather thin, 1 = very thin, and 0 means minimum fullness. The other scales work in similar ways. As you can see on the response form, it is possible to use decimals if you like.

The scales are 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 opposition to thin.
Spaciousness: The reproduction sounds open and spacious, in opposition to closed and shut up.
Brightness: The reproduction sounds bright, in opposition to dull and dark.
Softness/gentleness: The reproduction sounds soft and gentle, in opposition to sharp, hard, keen, and shrill.
Nearness: The sound seems to be close to you, in opposition to at a distance.
Loudness: The sound is loud, in opposition to soft (faint).
Fidelity: Judge how similar the reproduction is to the original sound. 10 = perfect fidelity, 9 = very good, 7 = rather good, and so on. (This scale is not used for the noise.)

There is a new response form for each reproduction. Mark your judgment on each scale by a straight vertical line. Do your ratings on each scale without looking at the other scales or earlier response forms. There are no right or wrong answers. It is solely your opinion about the sound that should be decisive.





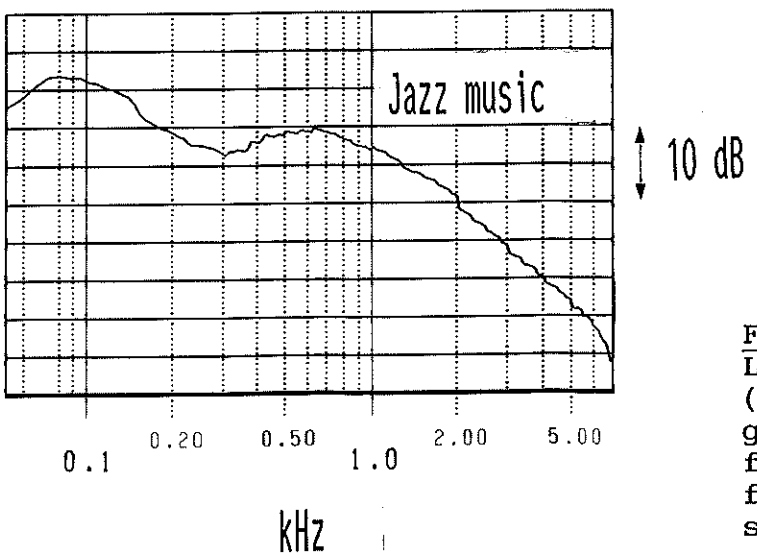


Figure 1.
Long Time Average Spectrum (LTAS) for the three programs. LTAS was calculated from the autocorrelation function of the program and smoothed across octaves.

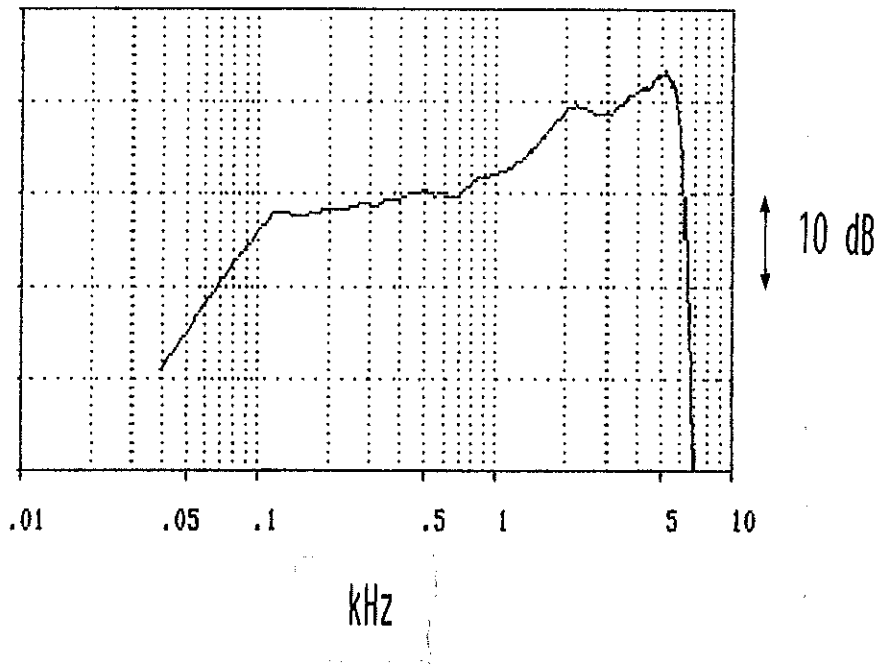


Figure 3. Frequency response of the earphone measured on a manikin (KEMAR) equipped with an ear simulator according to IEC 711.

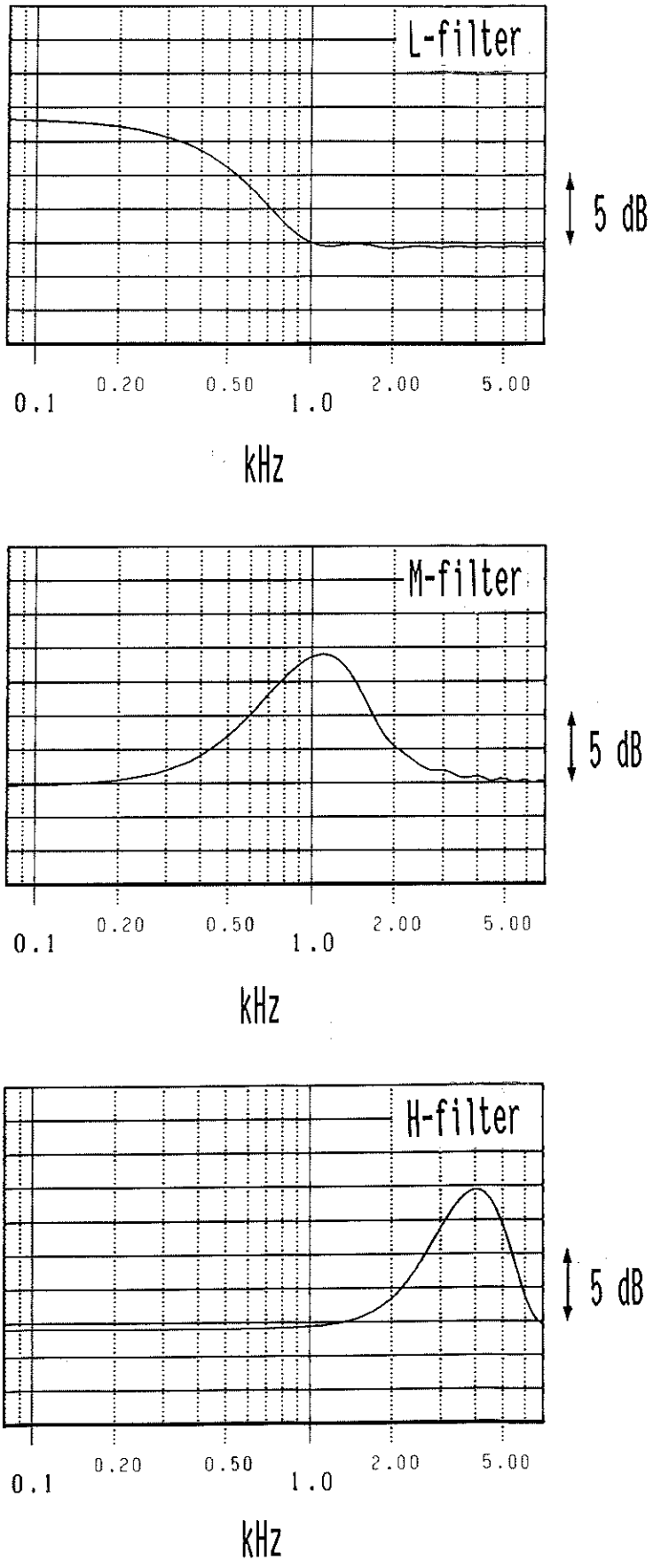


Figure 4.
Frequency responses of the three filters.

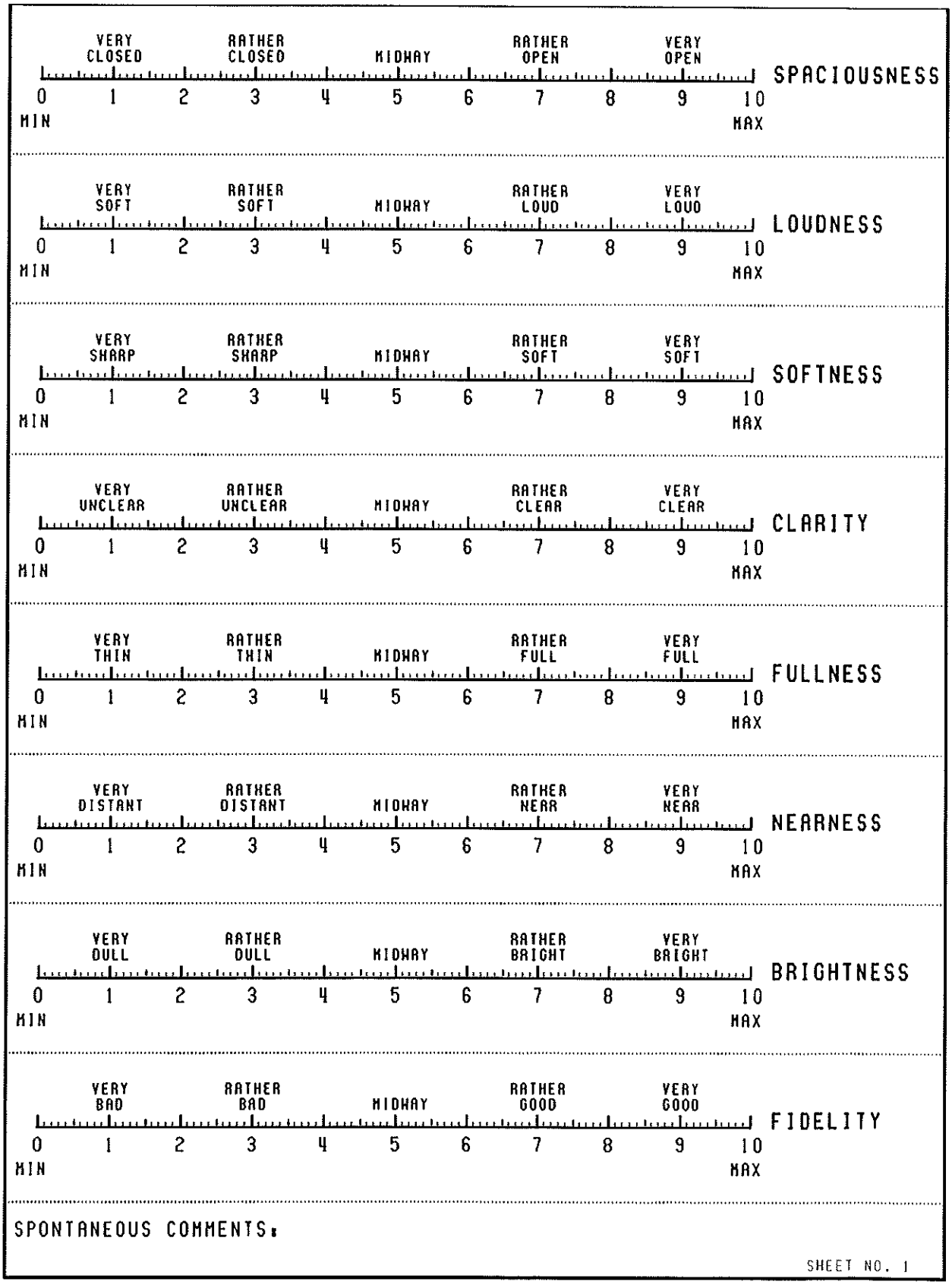


Figure 5. Example of the response form (translated from Swedish).

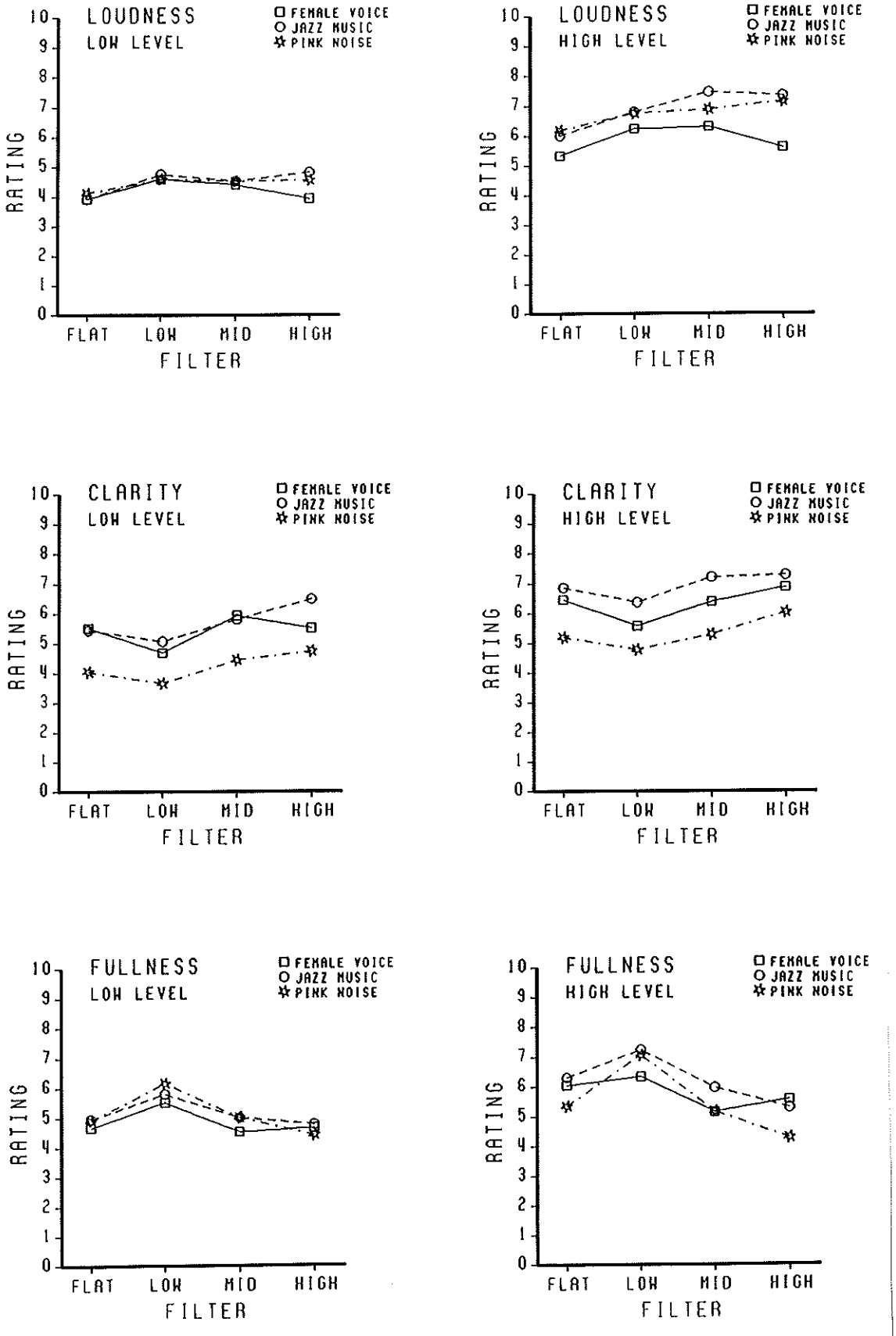
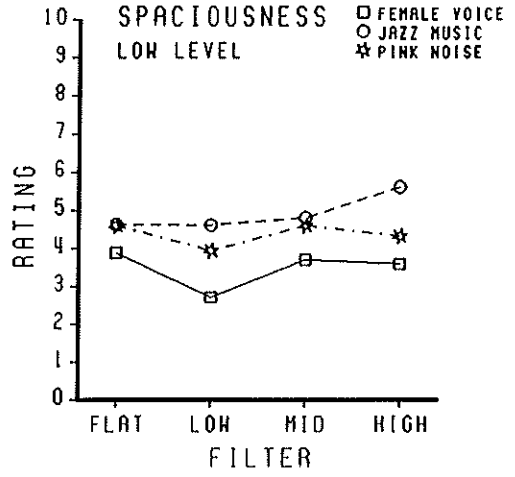
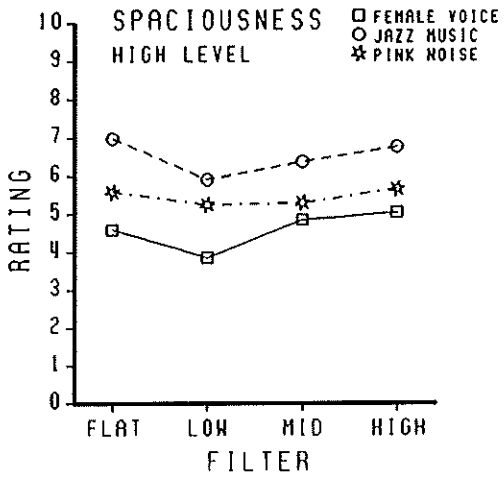
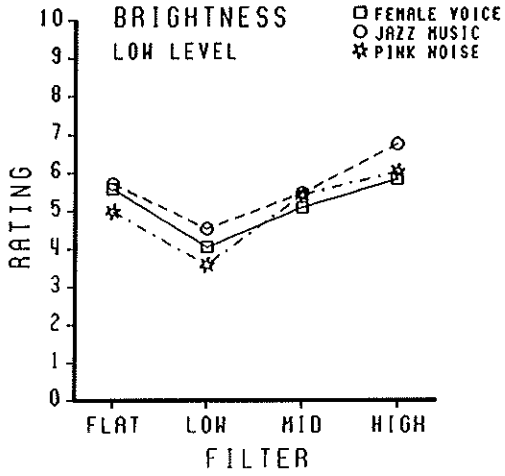
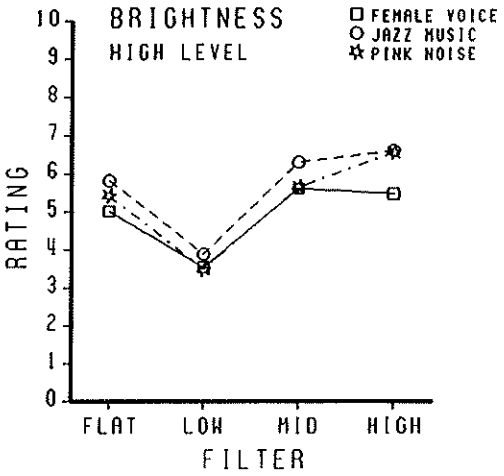


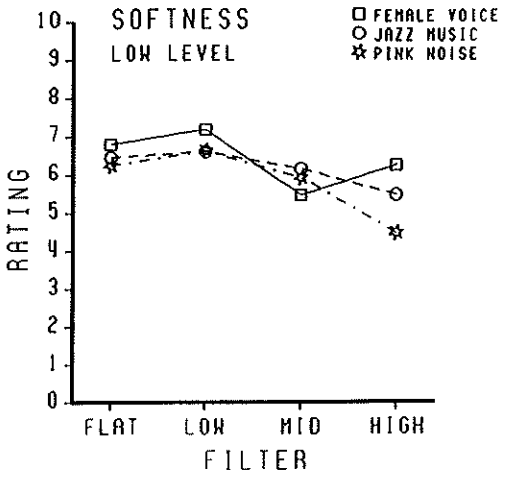
Figure 6. Average ratings across subjects in the different scales for the filtered reproductions at the two sound levels.











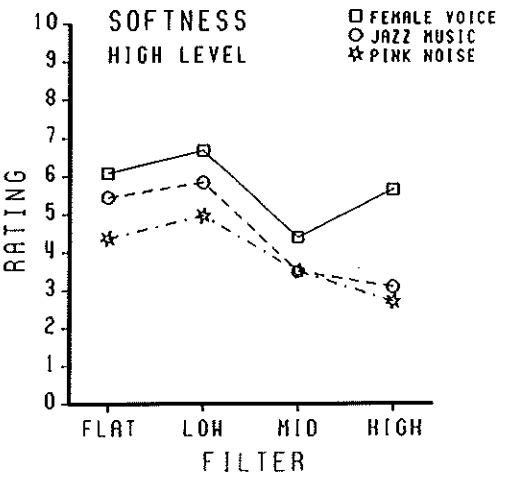


Figure 6. (Continued)

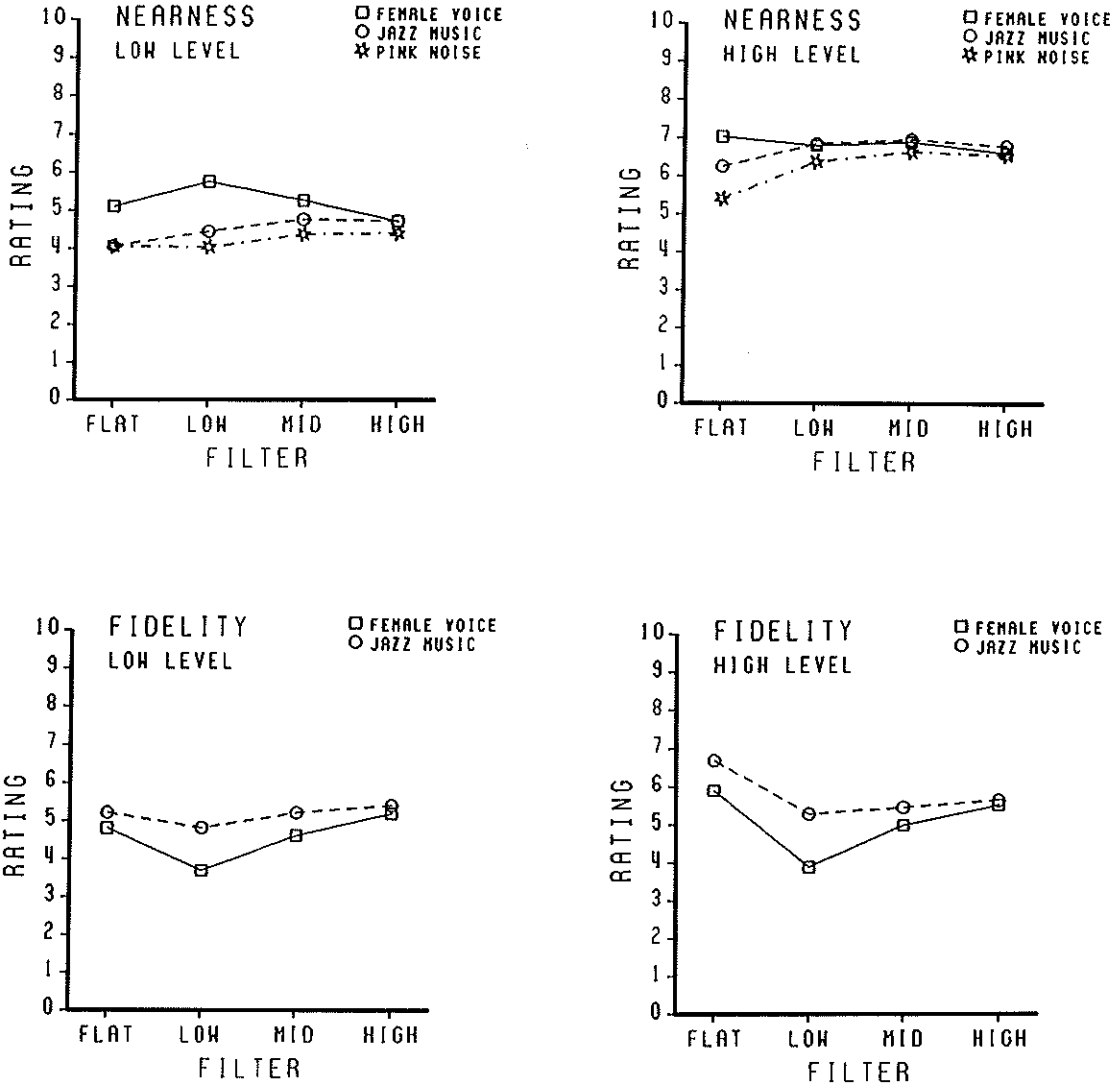


Figure 6. (Continued)