ISSN 0280-6819

KONGL CAROLINSKA MEDICO CHIRURGISKA INSTITUTET

TECHNICAL AUDIOLOGY

Report TA No. 114
December 1986

LOUDSPEAKER FREQUENCY RESPONSE AND PERCEIVED SOUND QUALITY: MEASUREMENTS IN LISTENING ROOM

Alf Gabrielsson, Björn Lindström and Ove Till

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

Report TA114
December 1986

LOUDSPEAKER FREQUENCY RESPONSE AND PERCEIVED SOUND QUALITY: MEASUREMENTS IN LISTENING ROOM

Alf Gabrielsson, Björn Lindström and Ove Till

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

ISSN 0280-6819
Report TA114
December 1986

LOUDSPEAKER FREQUENCY RESPONSE AND PERCEIVED SOUND QUALITY: MEASUREMENTS IN LISTENING ROOM

Alf Gabrielsson, Björn Lindström and Ove Till

ABSTRACT

The purpose of this investigation was to study the relationship between the perceived sound quality of 18 high fidelity loudspeakers (Gabrielsson & Lindström, 1985) and their frequency response in the listening room. Noise was fed through the loudspeakers and recorded by an omnidirectional microphone in the listening position. After A/D conversion the frequency response was calculated from the crosscorrelation function between the recorded noise and a reference noise. In a following step the frequency response was combined with the long time average spectrum of the music programs to approximate the stimulus reaching the listener's ears. Data from earlier listening tests were used to formulate hypotheses concerning the relationship between the frequency response and each of the perceptual variables clarity, fullness, spaciousness, brightness vs. dullness, softness vs. sharpness, absence of extraneous sounds, and an overall evaluation in terms of fidelity. Practically all hypotheses were confirmed by the present data.

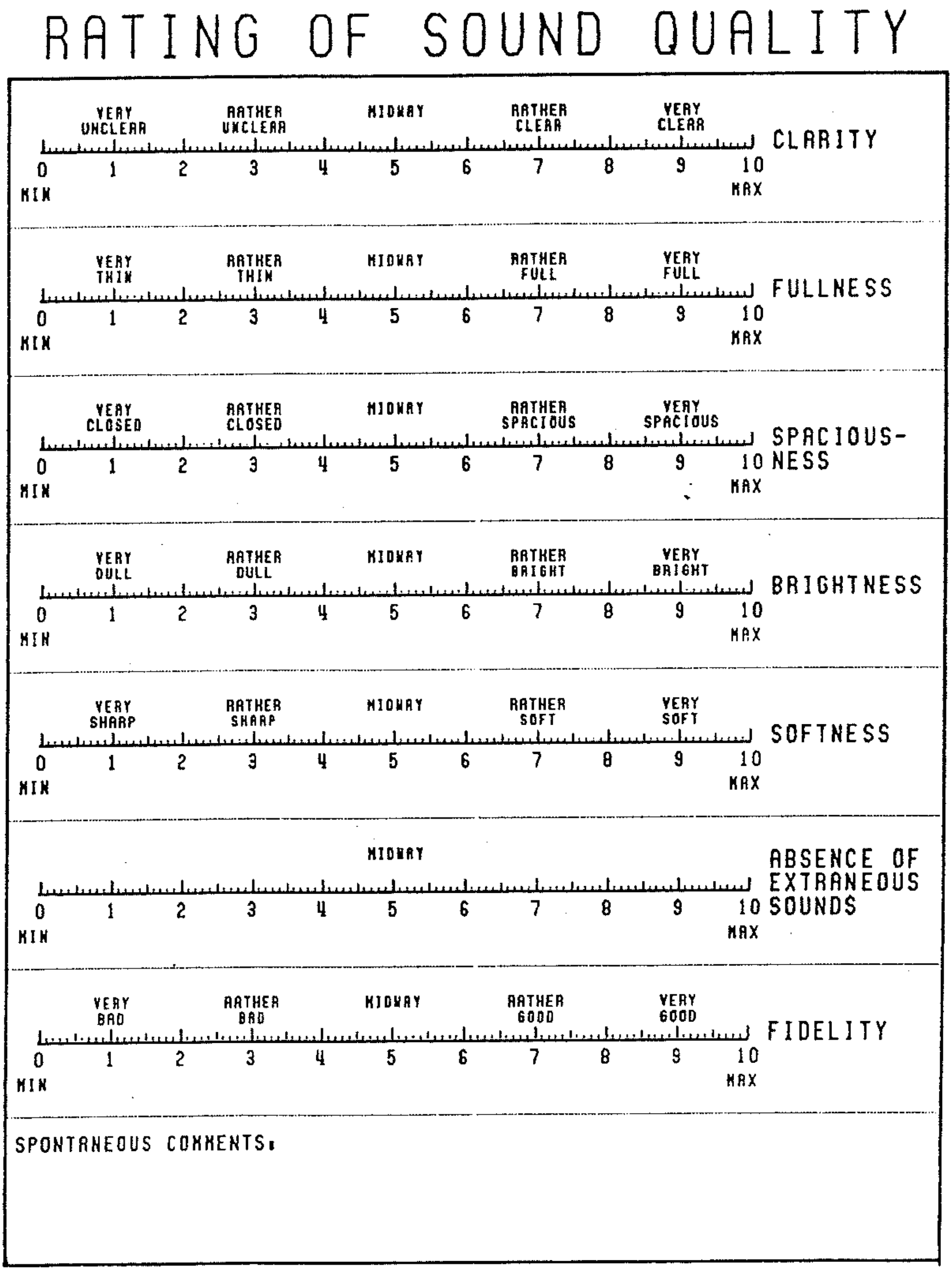
This work was supported by the Bank of Sweden Tercentenary Foundation and the Swedish Ministry of Health and Social Affairs, Commission for Social Research.

1. INTRODUCTION

Investigations on the perceived sound quality (PSQ) of loudspeakers, headphones, and hearing aids have shown that PSQ is composed by a number of separate perceptual dimensions: clarity, fullness, brightness vs. dullness, sharpness/hardness vs. softness, spaciousness, nearness, loudness, and (absence of) extraneous sounds (Gabrielsson, Rosenberg & Sjögren, 1974; Gabrielsson 1979, Gabrielsson & Sjögren 1979a, 1979b). These dimensions have been successfully used for perceptual description of loudspeakers (Gabrielsson, Frykholm & Lindström, 1979; Gabrielsson & Lindström, 1981; Gabrielsson, Lindström & Elger, 1983; Gabrielsson & Lindström, 1985) and for investigation of the effects of different frequency responses on perceived sound quality (Gabrielsson, Schenkman & Hagerman, 1985).

The relationships between the physical characteristics of the sound reproducing systems and the perceptual dimensions have been studied in various ways. In most of the experiments the selected loudspeakers (headphones, hearing aids) were judged in listening tests, and the results of the judgments were studied to see how they could be related to various physical properties of the systems. In some experiments the physical properties were deliberately manipulated by the experimenter (e.g., varying the frequency response or the amount of nonlinear distortion) in order to study the effects of these manipulations on the various perceptual dimensions. The results from many experiments were summarized by Gabrielsson & Sjögren (1979a), and they have also been mainly confirmed in more recent investigations (e.g., Gabrielsson, Schenkman & Hagerman, 1985).

In order to get sizable effects in the perceptual dimensions we usually selected widely different loudspeakers (headphones, hearing aids) in these earlier experiments, and when manipulations were made they were also quite large (e.g., increasing or decreasing the frequency response by 10dB/octave). However, in a recent listening test on high fidelity loudspeakers (Gabrielsson, Lindström & Elger, 1983; Gabrielsson & Lindström, 1985) the selection was made by the National Swedish Board for Consumer Policies and resulted in a set of 18 (pairs of) loudspeakers, which were considered as representative for loudspeakers included in brand-name hi-fi packages in the low and middle price range at that time in Sweden. The sound reproduction of these loudspeakers was studied for eight different music programs, using 16 experienced listeners. They rated the reproductions with regard to clarity, fullness, spaciousness, brightness, softness, absence of extraneous sounds, and (overall) fidelity. All rating scales were graded from 10 to 0, see Figure 1. The results were given in terms of the average rating over listeners for each loudspeaker x program combination in each of the seven scales and appear in matrices in the above-mentioned reports.



RATING OF SOUND QUALITY

VERY UNCLEAR	RATHER UNCLEAR	MIDWAY	RATHER CLEAR	VERY CLEAR	CLARITY

0 1 2 3 4 5 6 7 8 9 10
MIN MAX

VERY THIN	RATHER THIN	MIDWAY	RATHER FULL	VERY FULL	FULLNESS

0 1 2 3 4 5 6 7 8 9 10
MIN MAX

VERY CLOSED	RATHER CLOSED	MIDWAY	RATHER SPACIOUS	VERY SPACIOUS	SPACIOUS-NESS

0 1 2 3 4 5 6 7 8 9 10
MIN MAX

VERY DULL	RATHER DULL	MIDWAY	RATHER BRIGHT	VERY BRIGHT	BRIGHTNESS

0 1 2 3 4 5 6 7 8 9 10
MIN MAX

VERY SHARP	RATHER SHARP	MIDWAY	RATHER SOFT	VERY SOFT	SOFTNESS

0 1 2 3 4 5 6 7 8 9 10
MIN MAX

MIDWAY	ABSENCE OF EXTRANEOUS SOUNDS

0 1 2 3 4 5 6 7 8 9 10
MIN MAX

VERY BAD	RATHER BAD	MIDWAY	RATHER GOOD	VERY GOOD	FIDELITY

0 1 2 3 4 5 6 7 8 9 10
MIN MAX

SPONTANEOUS COMMENTS:

Figure 1 The response form (translated).

Extensive physical measurements have been made to study the relationships between these loudspeakers' physical properties and the listeners' ratings. There is ample reason to believe that the frequency response of the speakers is one of the most important factors. However, the frequency response can be measured in several different ways, e.g., in an anechoic room, in a reverberation room, or in the actual listening room. Furthermore the measurements can be made in different positions and at different points of time.

From the standpoint of general perception psychology it is preferable to try to describe the stimulus as accurately as possible. When perceiving music reproduced by loudspeakers, the stimuli to the listener's auditory system are extremely complex and continuously changing, and the acoustical properties of the room will also influence the impression of the sound. To study this process in all details would require too much time and labor and result in a prohibitively large amount of data. On the other hand, restricting the measurements to the loudspeaker's transmission in an anechoic room or in a reverberation room is a severe limitation. A reasonable compromise could be to perform measurements of the frequency response of the loudspeaker in the actually used listening room and combine this response with the long time average spectrum of the respective piece of music. This should give a reasonably adequate description of the stimulus reaching the listener's ears.

The basic idea is thus to try to describe the frequency response of the loudspeaker in the listening room and the spectral energy distribution of the stimulus in the listening position, and to relate these physical measures to the results of the listeners' ratings. We will refer to our earlier results concerning the psychophysical relations (e.g., those in Gabrielsson & Sjögren, 1979a) and use them as working hypotheses for the present comparisons. These results were not obtained by measurements in actual listening rooms, but the frequency responses were measured in reverberation rooms (for loudspeakers) or by various couplers (for headphones and hearing aids). However, the differences among the selected systems, or the manipulations undertaken with them, were usually so large (cf. above), that the psychophysical relations obtained in this way should be valid even if another method for measurement of the frequency response is used.

The hypotheses will be stated later in section 3.1.

2. METHODS FOR PHYSICAL MEASUREMENTS

The listening room was described in detail in Gabrielsson, Elger & Lindström (1983) and Gabrielsson & Lindström (1985). It was chosen and arranged in accordance with the recommendations in the IEC report on listening tests on loudspeakers (IEC Publication 268-13).

The room had the following dimensions: length 705 cm, width 465 cm and height 278 cm. The floor was carpeted, the ceiling was reflecting. Various absorbents (mineral wool) and reflecting surfaces were placed in different positions to affect the reverberation and diffusivity. The reverberation time was .355 sec in the frequency range 250-8000 Hz, and the deviations in different third octaves were within $\pm$ 20% of the mean. The background noise level was below 35 dBA.

The loudspeakers were placed in the rear part of the room, see Figure 2, according to the recommendations given by the manufacturers and IEC recommendations. They were concealed by an acoustically transparent curtain, 270 cm from the back wall.

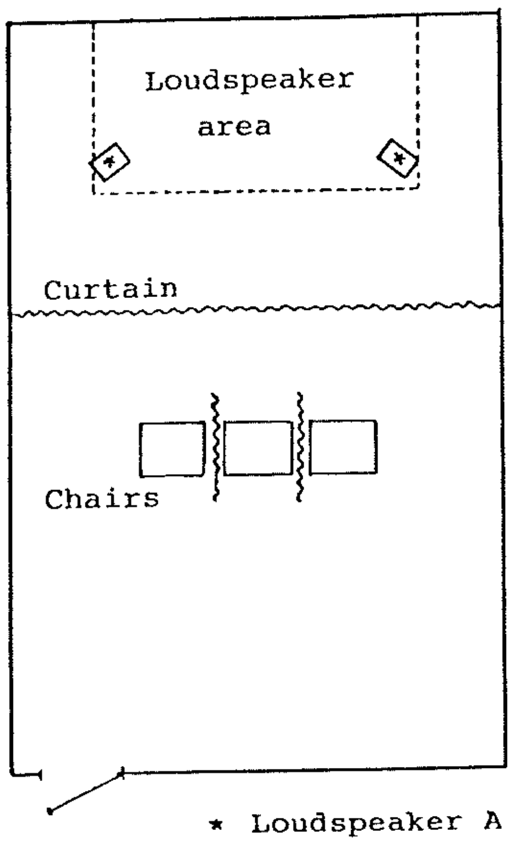


Figure 2. Listening room
Diffusors and absorbents like bookshelves and furniture are not shown in the figure.

Five or six pairs of loudspeakers were used in each listening session, and four sessions were needed to include all eighteen pairs of loudspeakers. One of them, loudspeaker A, was used in all four session to provide a common reference for comparisons between sessions (however, the listeners did not know about this). There were three listening positions at a distance of 130 cm from the curtain, see Figure 2.

The stimulus in the listening position was estimated from the long time average spectrum (LTAS) of the music program and the frequency transfer function of the left loudspeaker in the listening room. The analysis was divided into two parts:

i. The transmission channel described by the impulse response for the loudspeakers in the listening room (2.1), and

ii. A description of the music program (2.2).

2.1 Measurements of the impulse responses

As the listening room was located far away from the laboratory, registrations of noise were made on a stereo tape recorder for later analysis in the laboratory.

A general scheme of the measurement system is shown in Figure 3. Noise, 20 Hz - 20kHz, was fed through the equipment used in the listening test and recorded on one channel of the tape recorder. At the same time the noise was directly recorded on the other channel to provide a reference.

Recordings were made in the three listening positions of the left and the right speaker for all eighteen loudspeaker pairs. However, only the left loudspeakers recorded in the middle listening position were used in the later analysis, see 2.4. An omnidirectional microphone was used. These recordings were then A/D converted and processed in a mini computer. As seen in Figure 4a, the frequency response of the systems was calculated from the crosscorrelation function between the recorded loudspeaker noise and the reference noise after Hanning windowing (Oppenheim & Schafer, 1975).

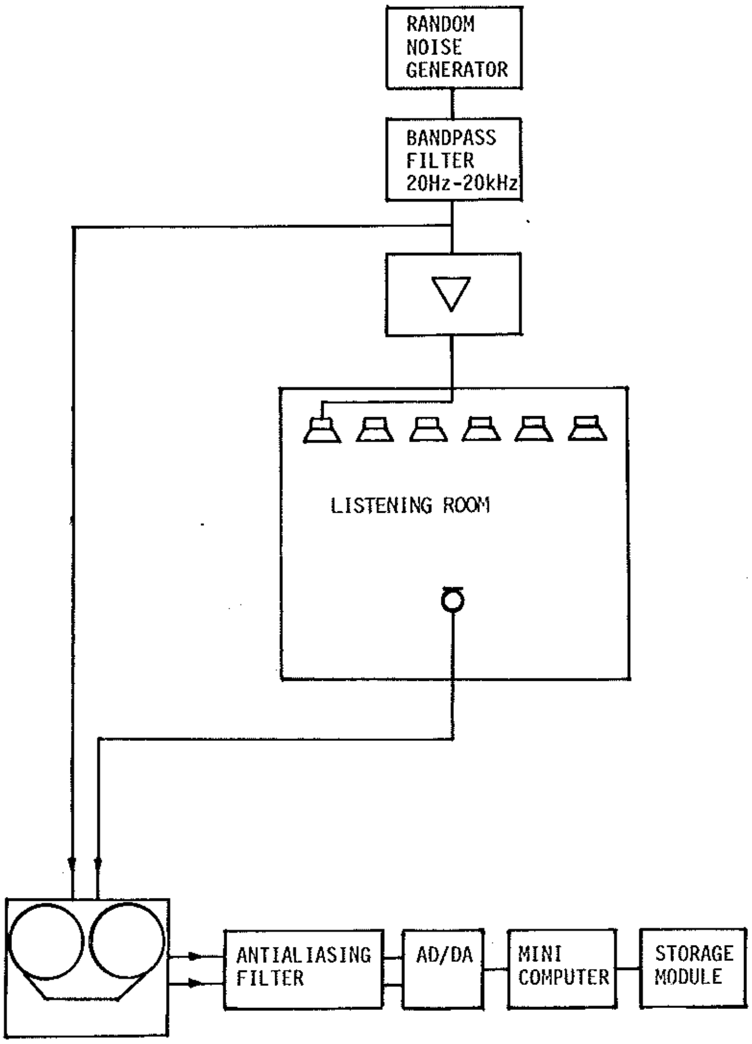


Figure 3. General scheme of the measurement system.

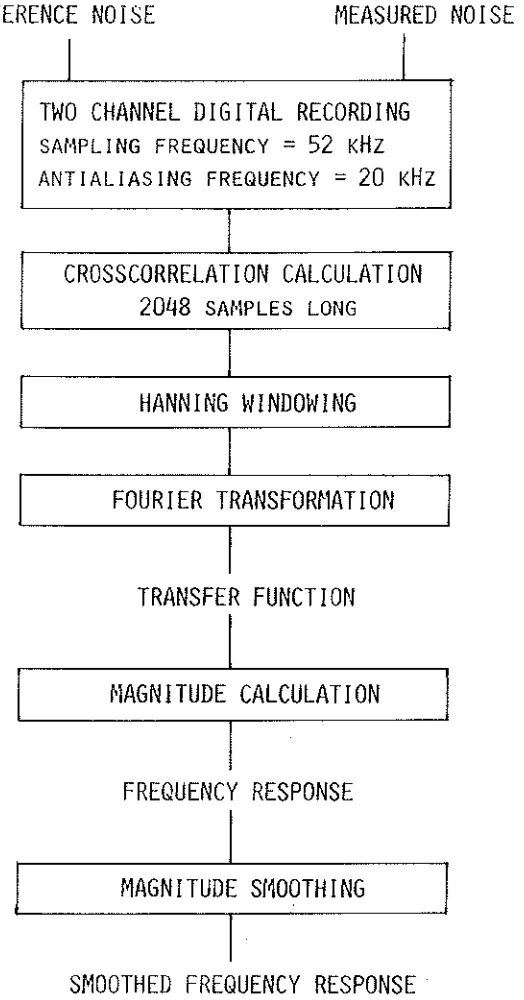


Figure 4a. Frequency response calculation.

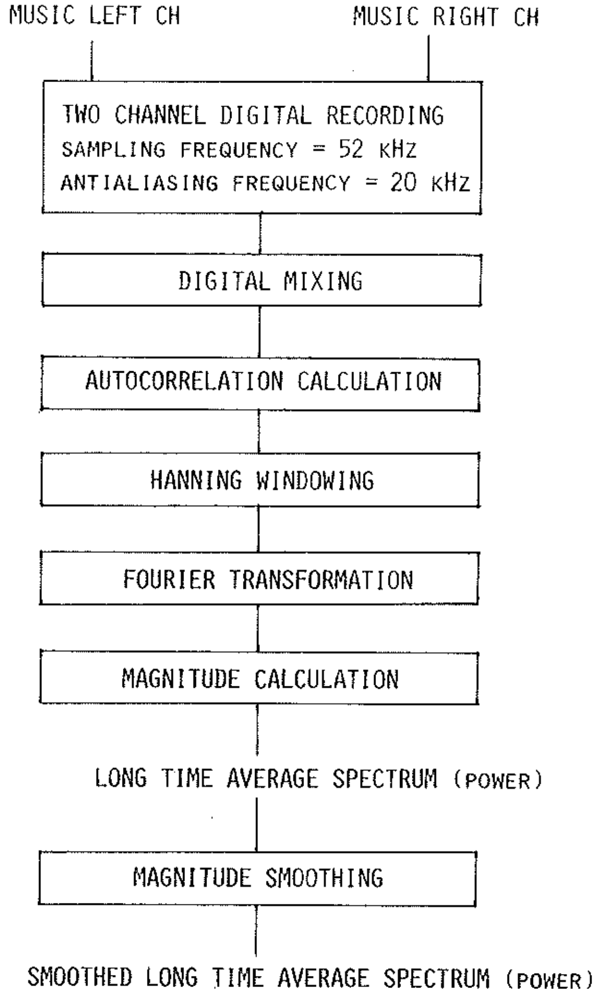


Figure 4b. LTAS calculation

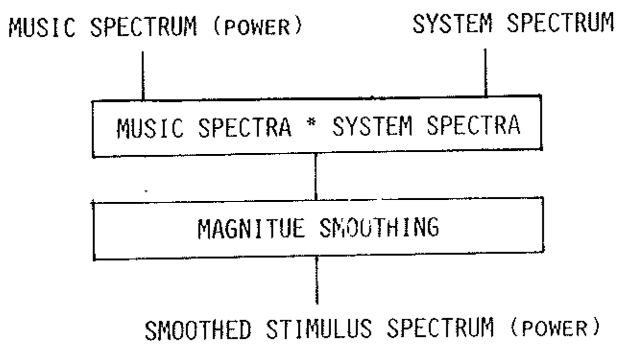


Figure 4c. Spectrum of the stimulus.

With the computer programs available the maximum usable length of the impulse response was 2048 samples. The frequency response was smoothed by using a moving average over the entire frequency range with a bandwidth of one octave. An example of a loudspeaker frequency response measured in the listening room is shown in Figure 5.

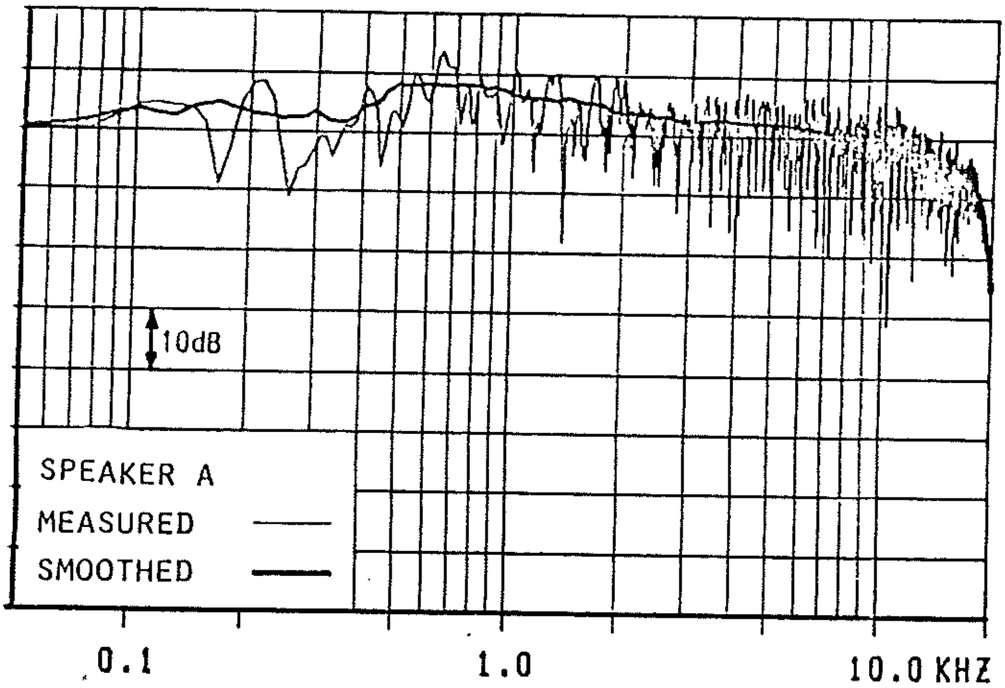


Figure 5. Example of frequency response measured in the listening room, unsmoothed (thin) and smoothed (thicker).

The smoothing of the spectrum was made as an energy average. This is the reason why the smoothed curve does not lie at the middle of the unsmoothed one when a dB scale is used. The smoothing is less accurate toward the ends of the frequency scale. At the upper end 16 kHz is the highest frequency that can have a full octave smoothing.

2.2 Long Time Average Spectrum of the music programs.

Four of the eight music programs were digitally recorded on a mini computer. The programs were tape copies from the master tapes or in one case a tape copy from a direct cut grammophone record. Each program lasted about one minute. The Long Time Average Spectrum (LTAS) was calculated from the autocorrelation function after a mixing to mono, see Figure 4b. The original and the octave smoothed spectra of the four programs are shown in Figure 6. The effects of the smoothing differ dependent on the shape of the original spectrum. Compare, for instance, the smoothing of the broad and uniform spectrum in the pop program with the smoothing of the narrower and more changing spectrum of the singer program.

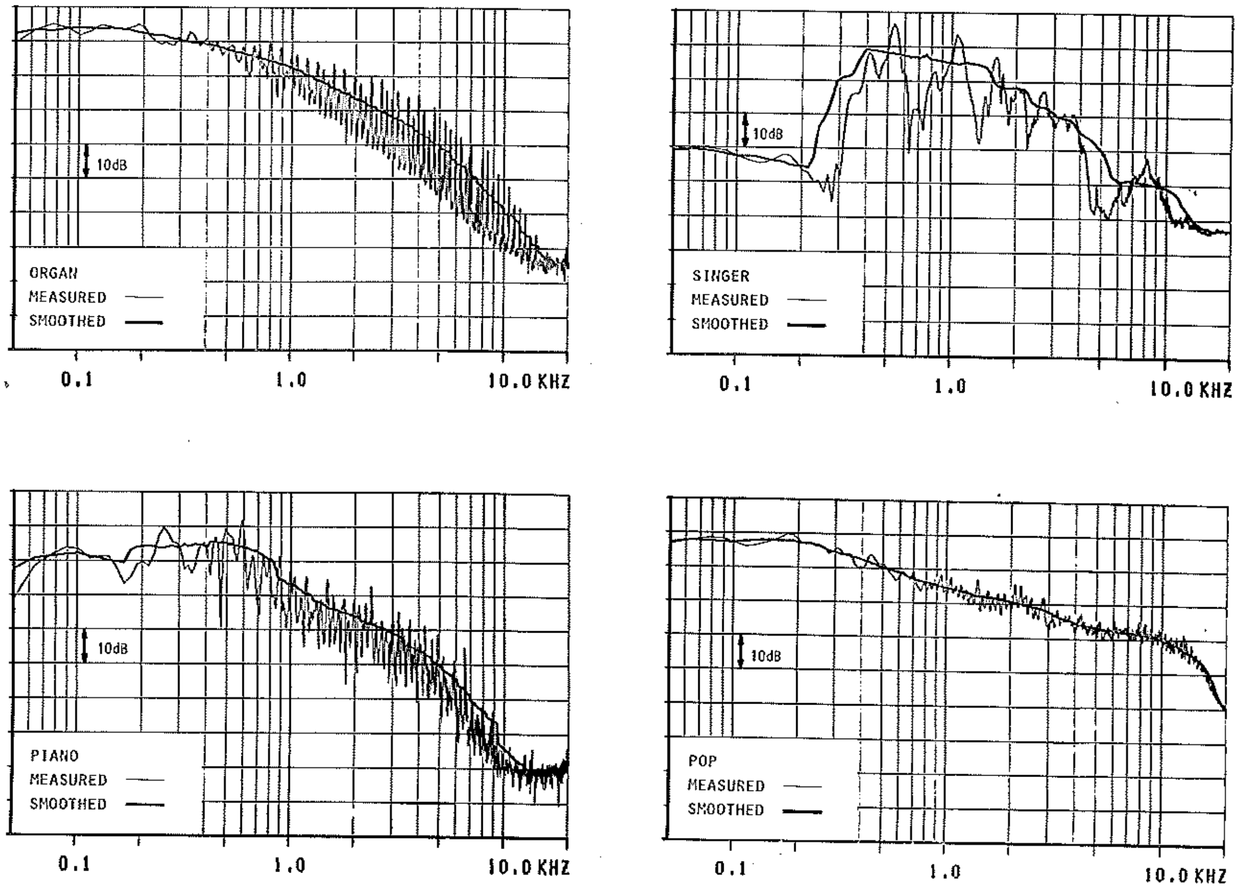


Figure 6. Original and octave smoothed Long Time Average Spectra of the four music programs.

2.3 Spectrum of the stimulus

As a measure of the actual stimulus in the listening position the product of squared frequency response of the loudspeakers and the music spectra was calculated, see Figure 4c. After this calculation, which was made with the unsmoothed data, an octave smoothing was performed. An example is shown in Figure 7, which shows the combination of the organ program (compare Figure 6) with loudspeaker A (compare Figure 5).

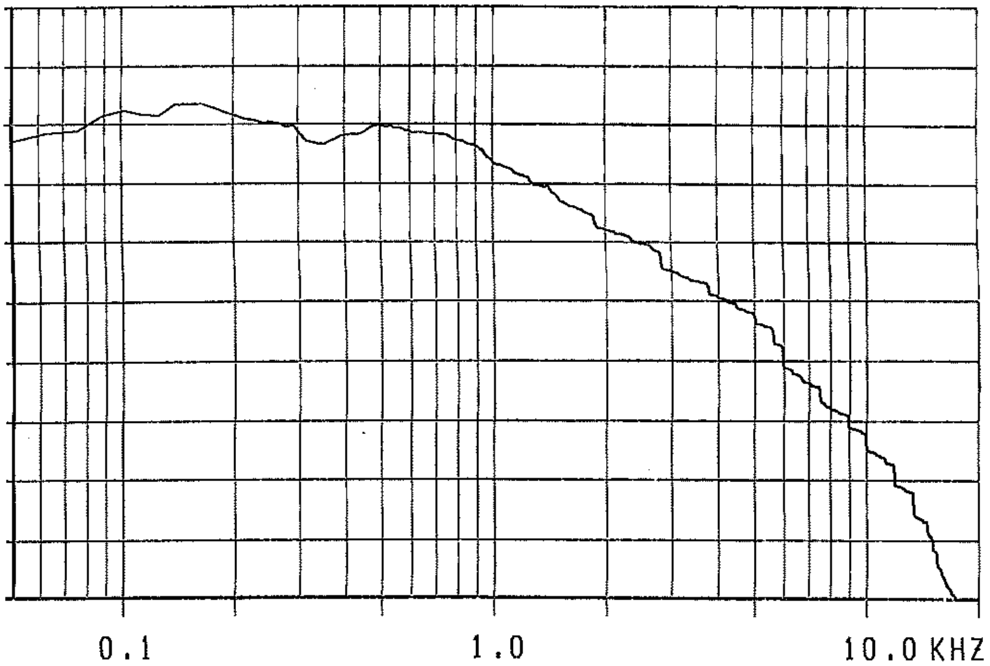


Figure 7. Example of calculated stimulus, combination of the organ program and loudspeaker A, after octave smoothing.

2.4 Assumptions and limitations in the estimation of the stimulus spectrum

To estimate the stimulus as a combination of the monophonic music spectrum and the frequency response of the loudspeakers measured with an omnidirectional microphone is obviously a simplification. It can to a large extent be seen as a way to reduce the amount of calculations and at the same time as a striving to separate the stimuli into as small parts as possible for special studies. Later on the analysis can be enlarged to incorporate more complex assumptions.
The use of an omnidirectional microphone, instead of an artificial head, shall be seen in the light of what was just mentioned and also against the background that we found an artificial head unsatisfactory in experiments on directional hearing.

To use a monophonic spectrum means that the stereo information in the signal is disregarded in the analysis. It may be expected that this will reduce the possibilities of interpreting the psychophysical relations regarding spaciousness, while the effects may be negligible for the other perceptual dimensions.

As stated in 2.1, only the signal from the left loudspeaker in the middle listening position was used in the analysis. It is assumed that the transfer function of the left and the right loudspeaker is very similar. The limitation to the middle listening position was motivated, besides the reduction of calculations, by the fact that it was not possible in the listening test to let each subject listen in all three different positions, but the listener had the same seat (middle, left, or right) during the whole test. Therefore the possible effects of different listening positions were confounded with the effects of different listeners. There was a certain tendency that the ratings given by the listeners in the mid-position were higher than for the listeners in the other two positions. However, this difference was not quite consistent and never approached any statistically significant diffarence.

The maximum length of the impulse response available from the calculations was 2048 samples at 52 kHz sampling rate, which is approximately 40 msec. The integration time of the hearing system can be discussed with respect to different types of perception. A number of authors (Lochner & Burger, 1964; Kuttruff, 1973) have investigated how the perception of speech is influenced by a single or multiple repetition of the direct sound. They found that sound arriving within 40 msec are almost fully integrated with the direct sound, while sound arriving between 40-100 msec gradually sounded more and more like separate sources or echoes. On the other hand studies of coloration (Atal, Schroeder, & Kutruff, 1962; Bilsen, 1968) have shown that a perceptual effect can be observed up to 30 msec with a maximum sensitivity around 2-5 msec. An impulse response of 40 msec therefore seems to be a reasonable compromise.

A critical point in the listening test and the physical analysis was the setting of listening levels. Before the listening test loudness matching was performed in order to make all loudspeakers sound equally loud when reproducing the same program (Gabrielsson & Lindström,1985). The correction values from this loudness matching were included in the calculation of the stimulus spectra (cf. 2.3 and Figure 9).

The use of long time average spectrum of the music programs has its subjective correspondence in the fact that the listener makes only one rating per scale, which is assumed to represent a kind of average over the whole program (about one minute). The programs were originally chosen to represent as homogeneous excerpts as possible.

Smoothing of the frequency response is of course advantageous with regard to general simplification. However, it is also reasonable to assume that the listener somehow performs a similar smoothing of the complex stimulus in perceiving and judging the sound quality. Rather than tracing all the ups and downs in the frequency response or in the spectral distribution (cf. Figures 5-6), the listener makes an integration (smoothing) into a limited number of frequency regions, e.g., sometimes only bass, midrange, and treble, or perhaps also with a subdivision within each of them, that is, lower and higher bass, lower and higher midrange, and lower and higher treble. Each of those categories encompasses at least an octave. The reduction to a small number of categories may be seen as an example of the well-known principle that our capacity for processing information is limited to '7 $\pm$ 2' categories (Miller, 1956). Although we don't know exactly how the listener makes his smoothing - and this probably varies depending on the characteristics of the music, the reproduction system, and the listening room - it seems that the smoothing performed in the analyses here may be a reasonable approximation.

3. RESULTS

3.1 Hypotheses

On the basis of the results obtained in many earlier listening tests summarized in Gabrielsson & Sjögren (1979a) and also in a recent investigation on the effects of different frequency responses (Gabrielsson, Schenkman & Hagerman, 1985), the relationships between the loudspeakers' frequency response and the perceptual dimensions are expected to be as follows:

Clarity is favored by a broad frequency range and by a certain (not too large) rise in the response at midfrequencies (about 250 - 1000 Hz) and the lower high frequencies (roughly 1-4 kHz). The more the emphasis (center of gravity) of the response is shifted toward lower frequencies, the worse clarity.

Fullness is likewise favored by a broad frequency range but with relatively more emphasis on lower frequencies.

Spaciousness seems to be favored by the same conditions as clarity. Narrow frequency range and/or marked resonances are negative factors for spaciousness as well as for clarity.

Brightness vs. dullness: Brightness increases when the emphasis of the response moves toward higher frequencies. Conversely, dullness increases when the emphasis of the response moves toward lower frequencies.

Softness (gentleness) vs. sharpness: Softness is favored by a certain emphasis on lower frequencies, while sharpness is associated with more or less steeply rising responses toward higher frequencies or marked resonance peaks at higher frequencies.

Absence of extraneous sounds (e.g., hiss and the like) is related to a (relatively) decreased response at very high frequencies (roughly above 5 kHz).

The hypotheses are admittedly not very precise in certain respects but their general meaning should be clear. They deal solely with the effects of the frequency response, all other things being equal. Some of the perceptual dimensions seem to depend on the frequency response in similar ways. However, this does not mean that they are different names for the same thing - they may still be phenomenologically different (Gabrielsson, 1979, p.167).

Beside the perceptual scales the listeners also made an overall evaluation in terms of fidelity. This may be considered as a weighted combination of the perceptual scales, and the relationship between fidelity and frequency response may accordingly be expected to reflect the relations between fidelity and the separate perceptual scales in proportion to their relative weights for achieving fidelity. These weights are rougly indicated by the correlations between the perceptual scales and fidelity (Gabrielsson & Lindström, 1985, Table 4), which are highest for clarity and spaciousness, followed by fullness, brightness, softness and absence of extraneous sounds in that order. The hypothesis is thus that fidelity is favored by a frequency response similar to that given above for clarity and spaciousness but somehow modified to reflect the influence of the other perceptual dimensions. Conversely, it is expected that fidelity will decrease for such frequency responses, which are unfavorable for one or more of the perceptual dimensions.

In the following the results of the measurements will be related to the judgments obtained in the listening test, first with regard to the loudspeakers' frequency response (3.2) and then with regard to the combination of music program and loudspeaker (3.3). To reduce the amount of calculations, this is made only for half of the complete listening test, which means that all 18 loudspeakers are included but only music programs Nos. 1-4 (organ, female singer, piano, "pop"), see 2.2 and Figure 6. These reproductions were judged by eight experienced and reliable listeners (Gabrielsson & Lindström, 1985).

3.2 Loudspeaker frequency responses and subjective judgments

The frequency responses of all 18 loudspeakers, measured as described in 2.1, appear in Figure 8. All of them have their distinct characteristics, even after octave smoothing as performed here, and it is not easy to classify them into a small number of categories. As an attempt to accomplish this, a factor analysis was made. First the average response curve over all 18 loudspeakers was calculated, and each speaker's difference against this average curve was obtained for 28 frequencies (those frequencies marked in Fig. 8 plus three others: 150, 1500, and 15000 Hz). The correlations between all speakers' "difference curves" were subjected to factor analysis (component analysis), and the solution was rotated according to the oblimin principle (Gorsuch, 1974). A four factor solution accounted for 92% of the total variance; however, three of those factors were bipolar. The corresponding frequency curves were calculated as the means of the frequency responses of those loudspeakers with factor loadings ≥ 0.84.

On the whole this analysis turned out to be of limited value. It provided some useful information about similarities between various frequency curves, which may be difficult to detect from solely visual inspection of so many curves. However, it was not possible to find enough consistent relations between the four factors and the perceptual dimensions.

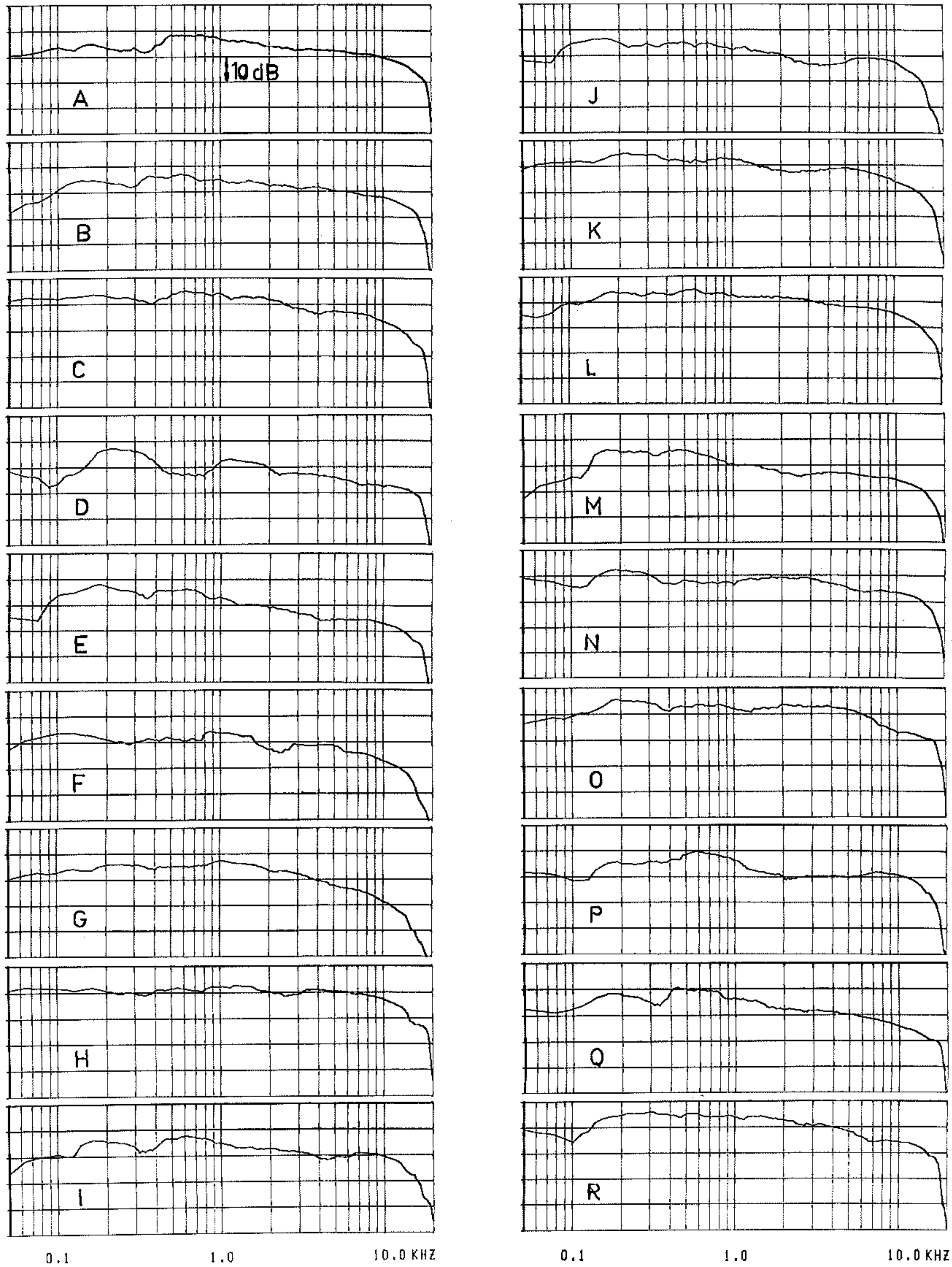


Figure 8. Frequency response of 18 loudspeakers after octave smoothing. The position along the vertical axis is arbitrary.

Rather than proceeding from similarities between the loudspeakers' frequency responses when trying to establish the relationships between them and the perceptual dimensions it was, not unexpectedly, more efficient to start from the perceptual side by picking out, for each perceptual scale, the loudspeakers which had been judged as the most different and look at the differences between their frequency responses. This was made separately for each of the four experimental sessions in the listening test to ensure that possible differences in judgments between the sessions would not affect the results. The first session included five loudspeakers (A,B,C,D,E), the second also five (A,F,G,H,I), the third session included six speakers (A,J,K,L,M,N) and the fourth included five (A,O,P,Q,R). Table 1 shows the loudspeakers which were rated highest and lowest, on average over programs Nos. 1-4, in each scale and session. For instance, in clarity the best speaker in the first session was A, the worst was E. In a few cases there was no statistically significant difference between the highest and lowest rated loudspeakers. These cases are indicated by a dash.

	Clarity	Fullness	Spaciousness	Brightness
Session 1	A - E	D - B	A - E	A - E
2	A - F	-	A - F	A - F
3	A - M	L - A	N - M	A - M
4	A - Q	R - P	A - O	A - Q

	Softness	Abs. extr. sounds	Fidelity
Session 1	E - C	-	A - B
2	I - H	G - H	A - F
3	-	-	L - M
4	R - A	Q - A	A - O

Table 1. The loudspeaker rated highest (left) and lowest (right), on average over programs Nos. 1-4, in each scale and listening session.

In clarity loudspeaker A was the best in all sessions in contrast to E, F, M, and Q, respectively. Loudspeaker A has a relatively flat response (within ±4 dB up to 10 kHz) with a maximum at 500-1000 Hz, see Fig. 8. Loudspeaker E has its maximum more "to the left", 100-700 Hz, and falls off more rapidly toward higher frequencies than A. Loudspeaker F is almost flat up to 1500 Hz with slight maxima at 800-1500 Hz and around 100 Hz. Loudspeaker M has a pronounced maximum at 150-600 Hz, and loudspeaker Q has something similar within 150-800 Hz. Thus all of the worst speakers have maxima more toward lower frequencies than loudspeaker A, and above the respective maxima the response decreases toward higher frequencies more rapidly than what happens for A. On the whole these results agree with the hypothesis regarding clarity. Loudspeaker A has the broadest frequency range in the sense that its response is the "flattest" considered over the whole frequency range (50-20000 Hz), and it has its maximum response at middle to high frequencies, while lower frequencies are somewhat suppressed.

With regard to fullness the extremes in the first session were loudspeakers D (best) and B (worst). In fact B has the flatter response but drops off below 100 Hz. D has a pronounced maximum at about 150-350 Hz (almost 10 dB higher than neighbouring regions) and decreases more rapidly toward higher frequencies than B. In the third session the best was loudspeaker L, while A was worst. L has the flatter response, and its maximum extends down to about 150 Hz, thus lower than for A. Something similar is seen in the comparison between R (best) and P (worst) in the fourth session: R has a rather flat maximum 150-1000 Hz, while the maximum response in P is around 500-800 Hz. These results are in partial agreement with the hypothesis concerning fullness in so far that an emphasis on relatively lower frequencies leads to more fullness.

In spaciousness loudspeaker A was best in three of the four sessions. The comparisons A-E and A-F were already made under clarity above, and the comparison between loudspeakers A and O (fourth session) reveals that O, like E and F, has more emphasis on lower frequencies than A. The marked drop from 6 to 10 kHz in loudspeaker O may also be noted. The comparison between N (best) and M (worst) in the third session shows that N has a much flatter response and relatively more at higher frequencies than M, which has a pronounced maximum at about 150-600 Hz. On the whole these results agree with those stated in the hypothesis regarding spaciousness.

In brightness the result is that loudspeaker A was throughout the brightest, while E, F, M, and Q were the dullest in the respective sessions. This is exactly the same loudspeakers that were contrasted in the clarity scale. As said under clarity, loudspeakers E, F, M, and Q all have maxima toward lower frequencies than loudspeaker A, and their response toward higher frequencies falls off more rapidly than what holds for A. These facts are in agreement with the hypothesis concerning brightness vs. dullness.

With regard to softness (gentleness) loudspeaker E was softest in the first session in contrast to loudspeaker C. It is obvious that loudspeaker E has relatively more emphasis on the lower frequencies (although not below 100 Hz) than loudspeaker C. In the second session loudspeaker I was softest and loudspeaker H least soft. Loudspeaker I has its emphasis at about 150-1000 Hz, while H has a practically flat response all the way up to 10 kHz. (Loudspeaker H has the flattest response of all the loudspeakers. It is interesting to note that it is not considered to be among the best loudspeakers.) In the last session loudspeaker R was judged as the softest in contrast to loudspeaker A. R has its maximum response relatively more toward lower frequencies than loudspeaker A and also drops off more at higher frequencies than A. Thus all the softer loudspeakers have relatively more emphasis on lower frequencies and less on higher frequencies, which supports the hypothesis.

Regarding absence of extraneous sounds only two comparisons are made: loudspeaker G (best) versus loudspeaker H in the second session and loudspeaker Q versus loudspeaker A in the last session. According to the hypothesis, absence of extraneous sounds (like hissing) is related to decreased response at high frequencies (say above 5 kHz). Actually loudspeakers G and Q show a lower response in this region than loudspeakers H and A, respectively. The esential effect of this is that the tape noise in the recorded programs is relatively more suppressed.

In fidelity, finally, loudspeaker A was considered best in three of the four sessions but loudspeaker L in one (the third) session. The comparisons between loudspeakers A and F and between A and O were already made in connection with clarity and spaciousness. Looking at the contrast between loudspeakers A and B in the first session shows that B has its emphasis on somewhat lower frequencies than A but also drops off considerably below 100 Hz. With regard to loudspeakers L (best) and M in the third session, it is evident that L has the flatter response, while M has a considerable boost at about 150-600 Hz. All together the results from these comparisons are in line with the general hypothesis concerning the factors affecting perceived fidelity (cf. 3.1).

As said earlier, all results above apply to perceptual ratings of the loudspeakers on average over programs Nos. 1-4. However, most of them also hold for the perceptual ratings in average over programs Nos. 5-8 in the listening test (Gabrielsson & Lindström, 1985). One consistent exception was that for programs Nos. 5-8 loudspeakers A and G were contrasted in spaciousness, brightness, and fidelity in the second session instead of loudspeakers A vs. F with programs Nos. 1-4. The main difference between loudspeakers A and G is that the response of G falls off much more at higher frequencies than in loudspeaker A. This fact is also in good agreement with the hypothesis for spaciousness, brightness and fidelity.

Actually the stimuli for the listeners' judgments were the combinations of music programs and loudspeakers. In view of the complex spectrum of any music program, and the likewise complex frequency response of the reproduction system (including that of the listening room), it is not surprising that there were often significant interactions between programs and systems in the listening test. This means that the perceived difference between the loudspeakers somehow varied with the different programs. In extreme cases the difference may be in one direction for a certain program but in the opposite direction for another program. Table 2 shows the best and the worst loudspeaker in each scale separately for each program. It is noted that which loudspeaker is the best and the worst is fairly often the same for the organ and pop programs, that is, the two programs with the broadest frequency range. The situation is more or less different for the singer and piano programs. For instance, while loudspeaker A very often was considered as the best for the organ and pop programs, this is not so pronounced for the singer program, and with the piano program loudspeaker A was often judged inferior or even worst. Many other examples of interactions can be seen in Table 2, which should be compared with Table 1, and also with the complete presentation of data from the listening test (Gabrielsson & Lindström, 1985). In view of such interactions it is preferable to supplement the analyses above with an investigation of how the combination of a loudspeaker's frequency response and the LTAS of a music program affects the perceived sound quality.

		CLARITY				SOFTNESS			
		Organ	Singer	Piano	Pop	Organ	Singer	Piano	Pop
	1	A - B	A - C	B - E	A - B	E - D	E - A	B - C	E - B
Ses-	2	A - F	A - F	I - F	A - F	-	A - H	I - H	G - H
sion	3	A - M	N - M	N - M	A - J	-	K - M	M - A	A - J
	4	A - Q	R - P	-	A - Q	R - O	-	P - A	-

		FULLNESS				ABS. EXTR. SOUNDS			
	1	D - B	B - C	B - A	A - B	-	-	-	D - B
Ses-	2	H - I	-	I - A	A - F	-	G - H	G - H	-
sion	3	K - A	L - K	N - A	L - M	-	-	L - J	A - M
	4	R - P	R - P	Q - A	R - Q	R - P	Q - A	O - A	-

		SPACIOUSNESS				FIDELITY			
	1	A - B	-	B - E	A - E	A - B	B - E	B - D	A - B
Ses-	2	A - I	-	I - A	A - G	A - I	A - F	I - A	A - G
sion	3	-	A - M	N - A	A - J	L - M	L - M	N - M	L - M
	4	A - O	A - Q	P - O	A - O	A - O	R - P	-	A - Q

		BRIGHTNESS			
	1	A - D	-	A - E	A - D
Ses-	2	A - G	A - F	I - F	A - G
sion	3	A - K	N - M	A - M	A - J
	4	A - Q	A - Q	A - Q	A - O

Table 2. The loudspeaker rated highest (left) and lowest (right) for each program and listening session

3.3 Program x loudspeaker combination and subjective judgments

The two loudspeakers, which were rated highest and lowest in the respective perceptual scale, were picked out separately for each program and session, and their frequency response was combined with the LTAS of the respective program as described in 2.3, to approximate the spectrum reaching the listener's ears. There were in all 16 such cases (4 programs x 4 sessions) for each scale, see Table 2. For reasons of space only half of them, that is, eight cases per scale, are presented here in Figure 9. The selection has been made to include as many loudspeakers as possible and to be representative for all 16 cases, although a certain arbitrariness is unavoidable. For the scale concerning absence of extraneous sounds only four cases are given, since there were fewer significant differences in this scale. Identification of program and loudspeakers is given within each graph. The highest rated loudspeaker is designated by a solid curve, while the lowest rated loudspeaker is designated by a dotted curve. Of course, the general appearance of each curve is dominated by the shape of the LTAS for the respective program, compare Fig. 6. The pop program has the most uniform spectrum, followed by the organ program, the piano program and finally the female singer program, in which there is little energy below 300 Hz. The position along the vertical axis is comparable between the different programs and program x loudspeaker combinations. The sound level was highest for the pop and organ programs, followed by the piano and singer programs.

The hypotheses concerning the relations between physical and perceptual variables are analogous to those in 3.1. However, in this case the physical variable is expressed in terms of the distribution of spectral energy over the frequency range rather than as a frequency reponse. Considering its relation to each of the perceptual variables gives the following picture:

With regard to clarity it is apparent from the eight graphs concerning clarity in Fig. 9 that the combination of any of the programs with the worse loudspeaker (designated by the dotted curve) is characterized by much more energy at lower frequencies (below 250 Hz) and in most cases also in the lower part of the mid-frequencies region (up to 600-800 Hz) than what happens in the combination with the best speaker. The difference is sometimes as big as 10 dB. In the singer program, for which the contents below 300 Hz is unimportant, the worse alternative also means less energy above 1000-1500 Hz. Taken all together, these facts mean that the center of gravity of the spectral distribution for the combinations with the worse speakers lies relatively more toward lower frequencies than in the combinations including the best speakers. This result is in agreement with the hypothesis concerning clarity.

In fullness the results are not as unambiguous as for clarity. The graphs in Fig. 9 for fullness show that with the organ, singer and piano programs the best reproduction is characterized by more energy at lower and/or midhigh frequencies, but also, to a varying extent, at higher frequencies. For the pop program, however, the situation is rather the opposite. These results are thus only in partial agreement with the hypothesis concerning fullness.

With regard to spaciousness the corresponding graphs in Fig. 9 show that, in most cases, the worse alternative (dotted curve) means more energy at lower and midhigh frequencies, sometimes very marked as for loudspeaker E at the pop program, but also with an exception as loudspeaker A with the piano program. There are also examples of higher energy for the worse alternative at higher frequencies, such as for loudspeaker O with the piano program and loudspeaker G with the pop program. In the last-mentioned case, however, this is reversed above 5 kHz, which still is an important region in the pop program. Considered all together the results are mainly in accordance with the hypothesis concerning spaciousness.

For brightness vs. dullness the situation seems quite clear. In all cases the dullest sounding alternative (dotted curve) exhibits a more or less marked boost at lower and/or midhigh frequencies (the difference against the brightest alternative is sometimes more than 10 dB). There are also examples of reduced energy at higher frequencies for the dull reproduction. The results are in agreement with the hypothesis regarding brightness and dullness.

Regarding softness vs. sharpness the corresponding graphs in Fig. 9 indicate that the softest reproduction (solid curve) is characterized by more energy at low and/or midhigh frequencies and/or less energy at higher frequencies. In other words, the sharpest sounding reproduction has relatively more energy at higher frequencies and less energy at lower frequencies than the softest. An exception from this is the comparison between loudspeakers B and C with the piano program. On the whole the results agree with the hypothesis concerning softness vs. sharpness.

The four graphs in Fig. 9 regarding absence of extraneous sounds indicate that the worse reproduction (dotted curve) has more energy at very high frequencies, roughly above 5 kHz, which is in accordance with the hypothesis. However, there are also other differences between the two reproductions in each case, the meaning of which must presently be left open.

In fidelity, finally, the worse reproduction (dotted curve) is characterized by more energy at lower frequencies and/or in the midfrequency region than in the best reproduction. The difference can be very marked, e.g., in the comparison between loudspeakers B and E with the singer and between B and D with the piano program. In some cases there is also a tendency to lowered energy at higher frequencies (above 1 kHz), in addition to the boost at lower or mid-high frequencies, in the worse reproduction. One of them is the combination of loudspeaker M and the piano program, which shows a marked emphasis on the region about 150-700 Hz at the expense of both lower and higher frequencies. Another case worth mentioning is the combination of loudspeaker F and the singer program, which has an increase in the region 800-1500 Hz and also below 200 Hz in comparison with the better reproduction. Although the latter fact does not directly affect the reproduction of the singing voice, it considerably amplifies the low-frequent background rumble in the church, where the program was recorded. Considered all together, the above-mentioned circumstances - relatively more emphasis on lower frequencies and/or marked emphasis on a limited frequency region for the worse reproduction - are factors which negatively affect clarity and spaciousness, which in their turn are among the most important factors for fidelity (cf. 3.1). Other perceptual dimensions may be negatively affected as well. The results are in accordance with the hypothesis concerning fidelity.

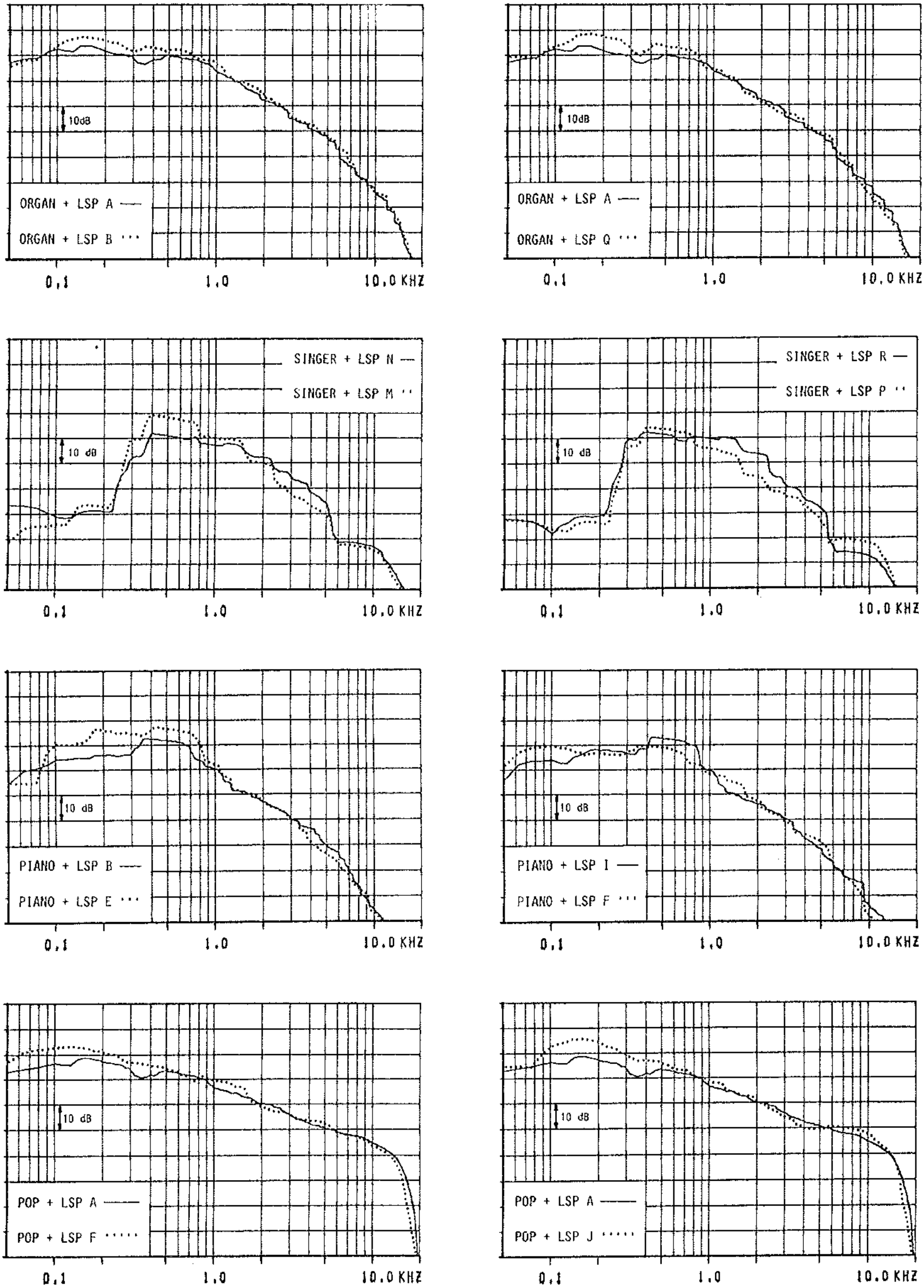


Figure 9. Spectra for the combination of a music program with the loudspeaker rated highest (solid curve) and rated lowest (dotted curve) in CLARITY.

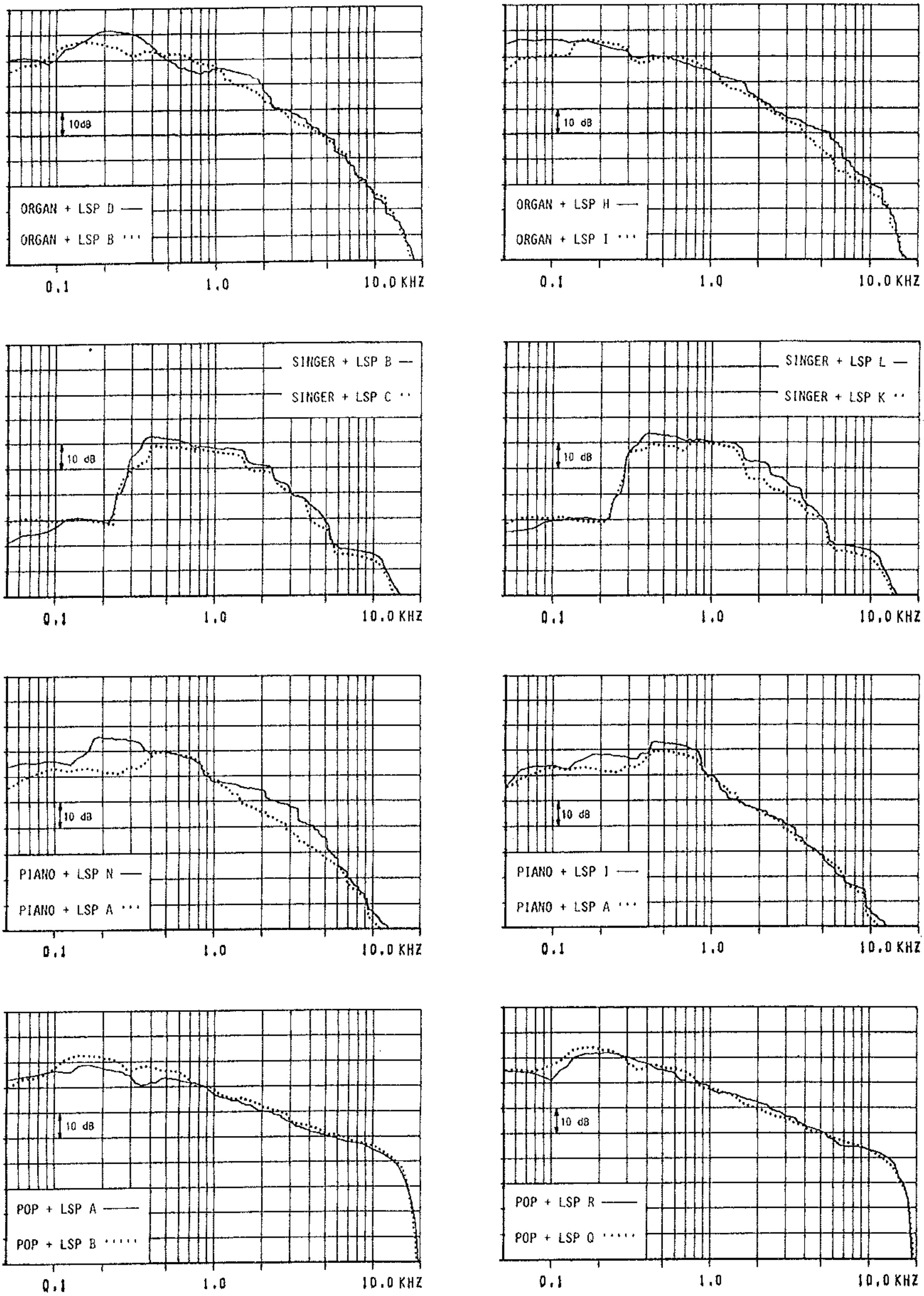


Figure 9 (continued). Spectra for the combination of a music program with the loudspeaker rated highest (solid curve) and rated lowest (dotted curve) in FULLNESS.

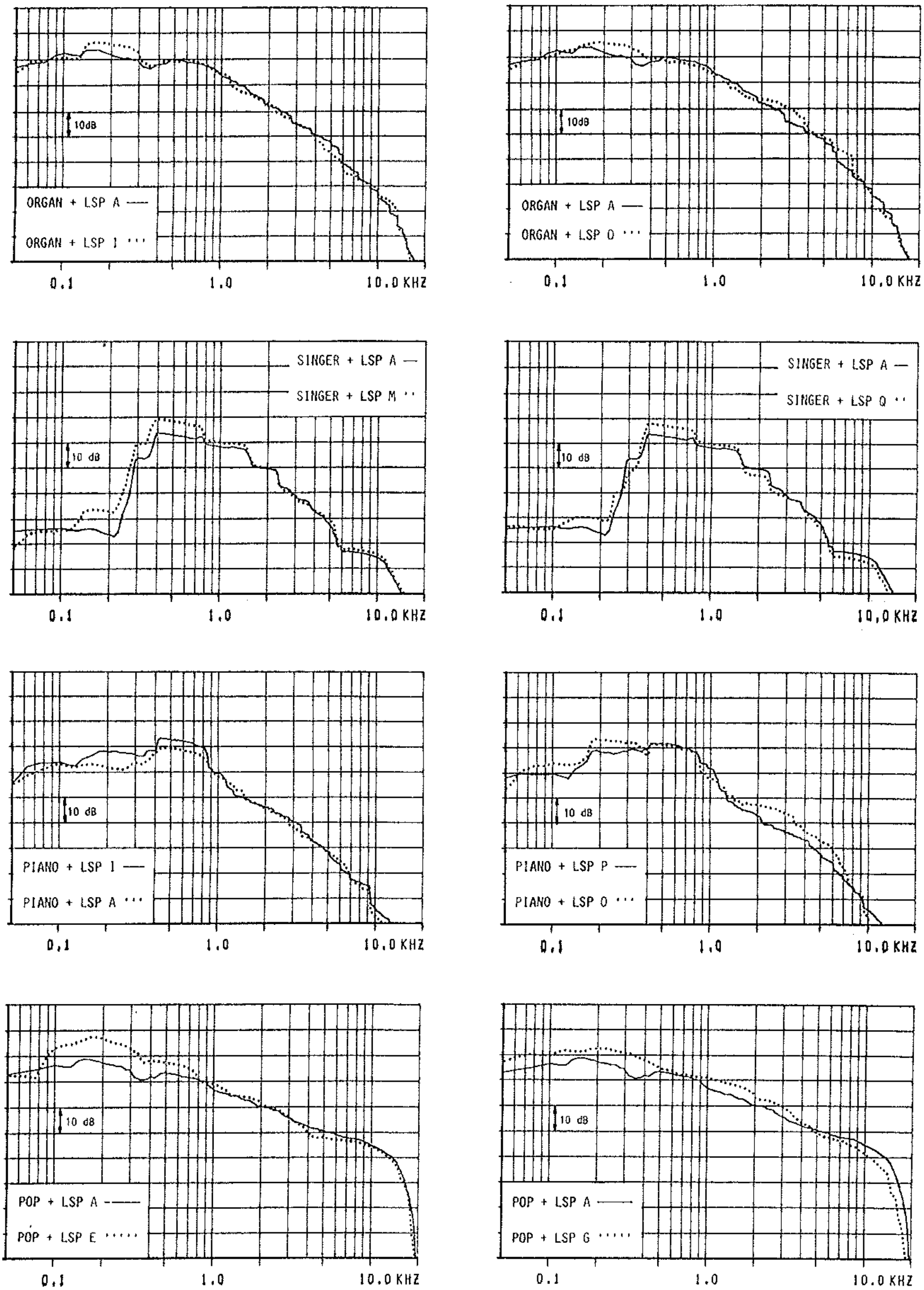


Figure 9 (continued). Spectra for the combination of a music program with the loudspeaker rated highest (solid curve) and rated lowest (dotted curve) in SPACIOUSNESS.

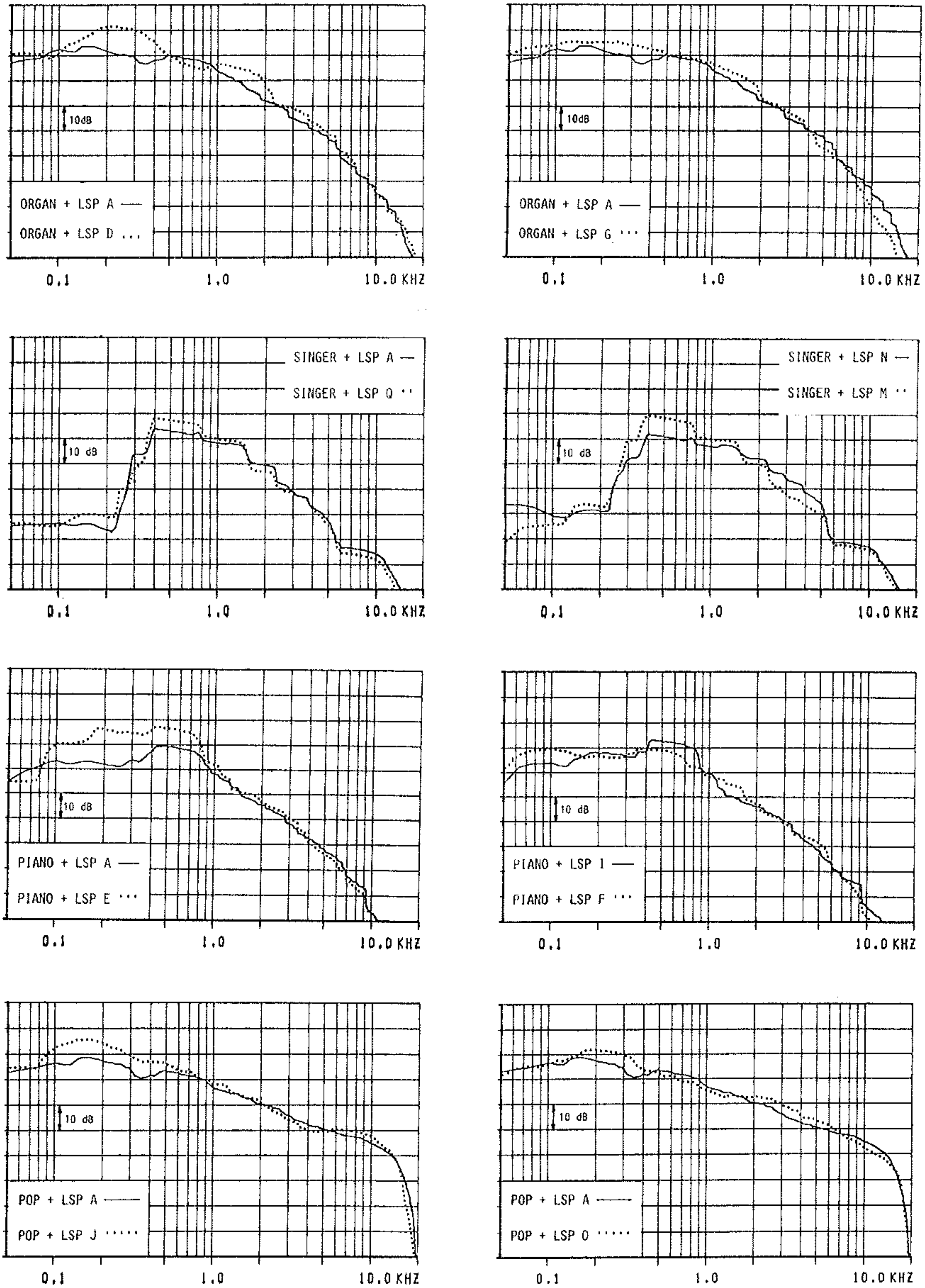


Figure 9 (continued). Spectra for the combination of a music program with the loudspeaker rated highest (solid curve) and rated lowest (dotted curve) in BRIGHTNESS.

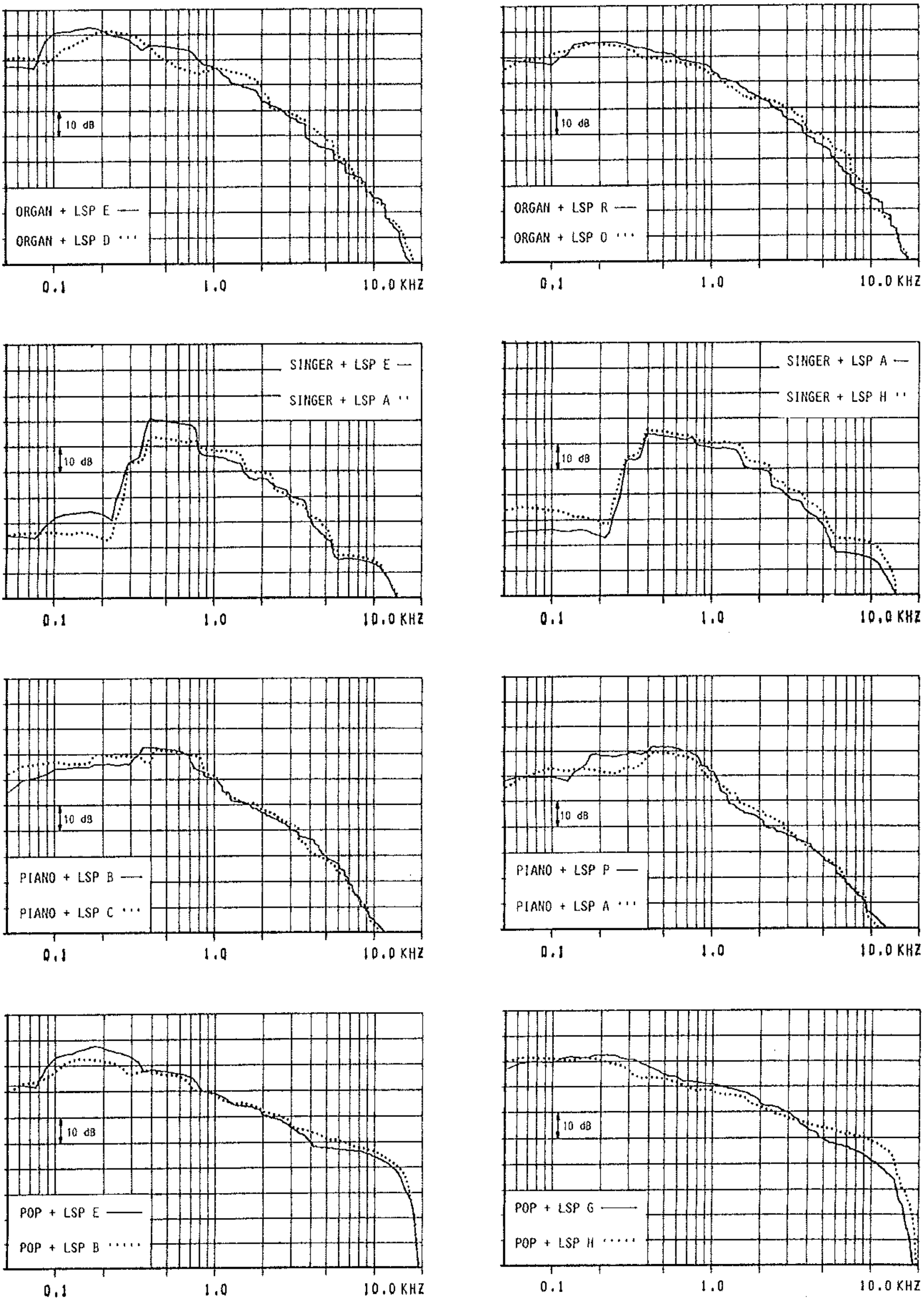


Figure 9 (continued). Spectra for the combination of a music program with the loudspeaker rated highest (solid curve) and rated lowest (dotted curve) in SOFTNESS.

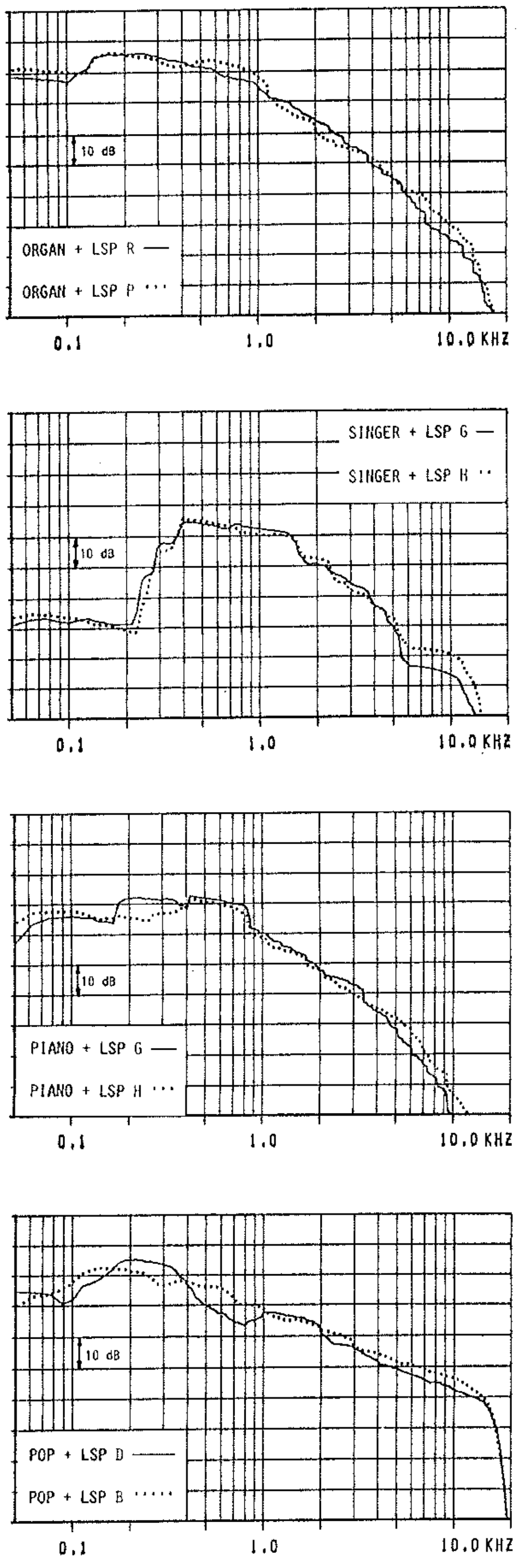


Figure 9 (continued). Spectra for the combination of a music program with the loudspeaker rated highest (solid curve) and rated lowest (dotted curve) in ABSENCE OF EXTRANEOUS SOUNDS.

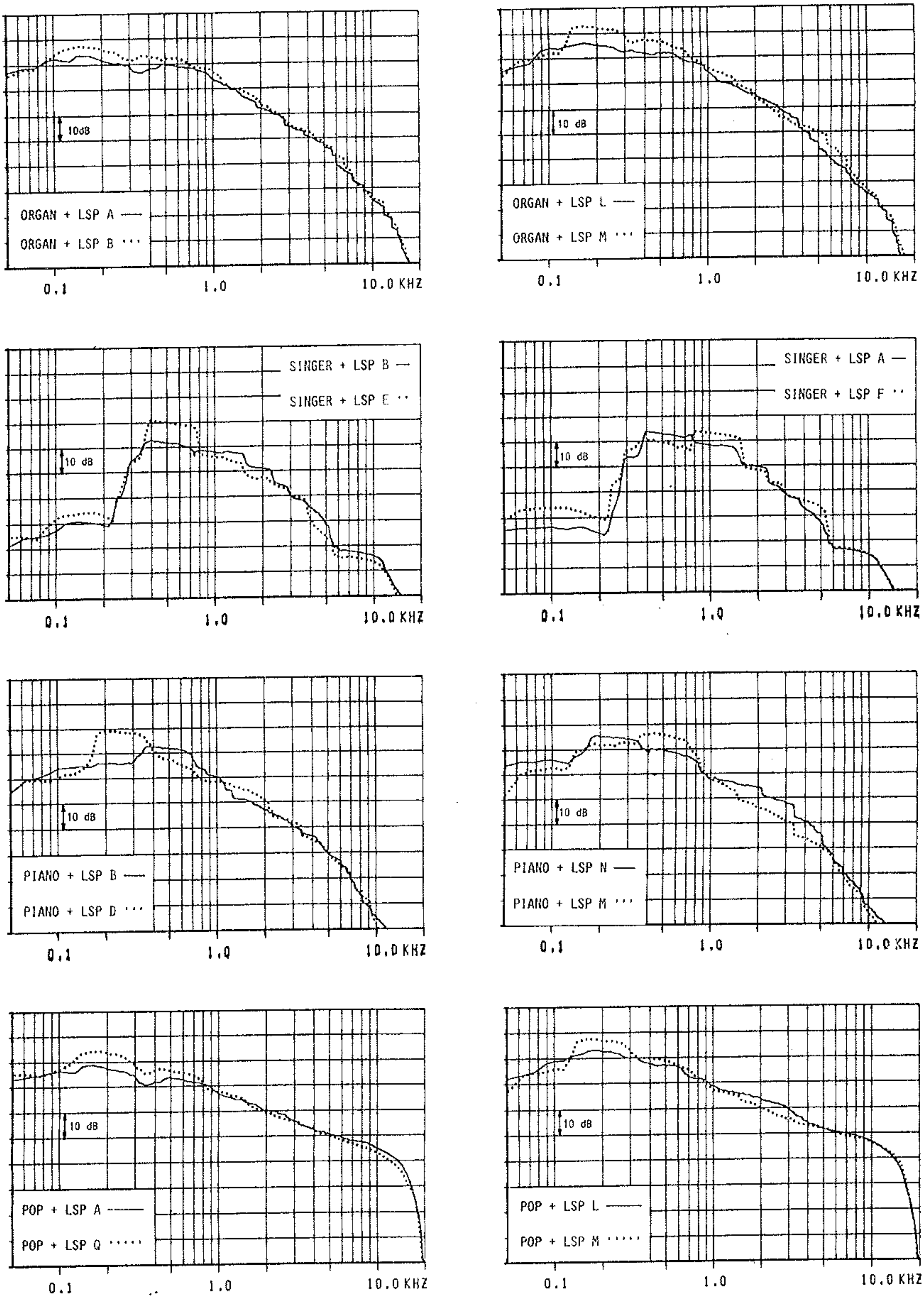


Figure 9 (continued). Spectra for the combination of a music program with the loudspeaker rated highest (solid curve) and rated lowest (dotted curve) in FIDELITY.

4. DISCUSSION

Although the estimation of the frequency response and of the stimulus spectrum was simplified in many respects (using omnidirectional microphone, octave smoothing, monophonic spectrum etc., cf. 2.4), the results are in good agreement with our findings from earlier listening tests, in which other measurement methods were used (cf. Introduction). The hypotheses concerning the relationships between the frequency response and the various perceptual scales were thus confirmed. There were some few exceptions, the interpretation of which is presently left open. All perceptual dimensions seem to be influenced by variations in the frequency response, even spaciousness, although the stereo information in the signal was discarded (cf. 2.4).

The hypotheses in 3.1 were explicitly stated as dealing only with the frequency response, all other things being equal. However, in this case we actually examine the hypotheses under circumstanses, in which the last condition - all other things being equal - is not fulfilled. The eighteen loudspeakers do not differ only with regard to the frequency response but in other factors as well, and it is thus possible that the listeners' ratings are influenced not only by the differences in frequency response but also by other factors. Definite conclusions concerning these questions have to await the results from further investigations.

The general impression from the evidence in Figures 8-9 is that the differences in frequency response and spectral distribution mainly belong to frequencies below, say, 1 kHz. Usally the worse alternatives have a more or less pronounced boost somewhere in this region. Besides making them sound duller, this also means more upward spread of masking, which has a negative effect on the perceived clarity, and thereby also a negative effect on fidelity.

In a forthcoming report a comparison will be made concerning the relationship between frequency and perceived sound quality, when the frequency response is measured in different ways: in the listening room as here, in an anechoic room, and in a reverberation room.

5. ACKNOWLEDGEMENTS

The authors want to express their gratitude to Gerhard Elger for much assistance in the physical measurements and to Björn Hagerman and Åke Olofsson for valuable discussions.

6. REFERENCES

ATAL, B.S., SCHROEDER, M.R., & KUTTRUFF, K.H. (1962). Perception of Coloration in Filtered Gaussian Noise - Short-Time Spectral Analysis by the Ear. _4th International Congress of Acoustics_, Copenhagen, Denmark, Paper H31.

BILSEN, F.A. (1968). _On the Interaction of Sound with Its Repetition_. PhD dissertation. Delft, The Netherlands: Uitegererij Waltman.

GABRIELSSON, A. (1979). Dimension analyses of perceived sound quality of sound-reproducing systems. _Scandinavian Journal of Psychology_, _20_, 159-169.

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. (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., 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., ROSENBERG, U., & 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., & HAGERMAN, B. (1985). The effects of different frequency responses on sound quality judgments and speech intelligibility. _Technical Audiology Reports_ No. 112. Stockholm: Karolinska Institute.

GABRIELSSON, A., & SJÖGREN, H. (1979a). Perceived sound quality of sound-reproducing systems. _Journal of the Acoustical Society of America_, _65_, 1019-1033.

GABRIELSSON, A., & SJÖGREN, H. (1979b). Perceived sound quality of hearing aids. _Scandinavian Audiology_, _8_, 159-169.

GORSUCH, R.L. (1974). _Factor Analysis_. Philadelphia: Saunders.

IEC REPORT, Publication 268-13. Listening Tests on Loudspeakers, Geneva 1985.

KUTTRUFF, K.H. (1973). Room Acoustics. London: Applied Science.

LOCHNER, J.P.A. & BURGER, J.F. (1964). The Influence of Reflections on Auditorium Acoustics. Journal of Sound and Vibration, 4,426-454.

MILLER, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

OPPENHEIM, A. & SCHAFER, R. (1975). Digital Signal Processing. Englewood Cliffs, New Jersey: Prentice Hall.