euronoise

Loudspeaker sound quality: comparison of assessment procedures

Vincent Koehl and Mathieu Paquier

LISyC EA 3883, 6 avenue Victor Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France
vincent.koehl@univ-brest.fr

In listening tests involving different loudspeakers and aimed at assessing the sound quality of these sound-reproducing systems, the level is generally adjusted to compensate for differences in sensitivity. The loudness sensation must be alike for each system under test. Because of the non-stationary nature of the musical signals used as test material in loudspeaker ratings, loudness assessment by using the current models (Zwicker, Moore...) remains slightly inaccurate. In practice, loudness is often equalized by ear by the experimenter. This study deals with the comparison of various test procedures. The first experiment was a paired comparison of loudspeakers where short-duration stimuli were presented to listeners for preference ratings. In the second experiment, the same listeners were allowed to switch, at any time, from one loudspeaker to another one so that the proposed stimuli were longer. In both experiments the loudness was equalized by the experimenter. However, under normal listening conditions, the listener is usually free to adjust by himself the reproduction level. At last, in a third experiment, the listeners had the opportunity, at any time, to not only switch from one system to another one, but also adjust the loudness of the stimuli.

1 Introduction

A subjective evaluation of loudspeakers is a difficult and time-consuming task. The parameters to be controlled are numerous and the results are often very context-dependent. Though the Audio Engineering Society and the International Electrotechnical Commission have both provided recommendations for loudspeakers listening tests [1, 2], till now no standardized technique has been adopted. Nevertheless, in sound quality evaluation, some existing comparison procedures appear as the most achievable and reliable. Clear guidelines for these tests are as follows: i) the loudspeakers under test are usually presented as pairs, which facilitates the comparison; ii) various sound events have to be tested; iii) loudness must be equalized over all of the loudspeakers involved in a session; iv) short stimuli must be preferred and v) the loudspeaker positions must be exchanged throughout the experiment to avoid positional effects.

All of these constraints make listening tests very far from a realistic listening situation. Under realistic conditions of comparison, a listener (audiophile, sound engineer…) usually listens to various types of music (generally rather long excerpts) and can even vary the reproduction level. Such an approach is totally different from the listening tests carried out within a laboratory.

This study was aimed at comparing three different assessment procedures, all based on paired comparisons. Four different loudspeakers were under test to determine whether their preference ratings are affected by the assessment procedure.

2 Paired comparison

Paired-comparison appears as an easy and reliable way to estimate loudspeaker sound quality. According to Toole [3], comparisons must be as quick as possible to ensure maximum discrimination and minimum variability in the judgments. Even when absolute judgments are desired, the loudspeakers are presented by pairs for maximum discrimination. This procedure is referred by IEC as paired ratings [2]. In listening tests involving paired comparison, the two stimuli are generally matched in loudness. A strong relationship between the perceived sound quality and the reproduction level was established in [4].

2.1 Reproduction level

Because of the influence of loudness on sound quality judgments, a common prerequisite is to check that the perceived reproduction level is alike for all of the loudspeakers under test. The current loudness models (Zwicker [5] or Moore [6]) have been acknowledged for stationary sounds, but not for the musical signals commonly used for loudspeaker comparison. Therefore, in loudspeakers listening tests the loudness is generally matched by ear by the experimenter himself or by some expert listeners.

This matching is one of the biggest differences between the listening tests carried out in laboratory and real-life comparisons. Therefore, it appeared interesting to design a test where the listener was free to set the reproduction level to his convenience and to compare the results to loudness-matched experiments.

2.2 Presentation method

In paired comparisons, the subject listens to either successive stimuli or alternate ones:

- The first method is the so-called A-B comparison, where the excerpt to be listened to is at first presented over the loudspeaker system, A, and then over the other one, B. They have to be compared one just after the other, because of the short human auditory memory [7]. This procedure appears as the most exact way to compare the same excerpt. But, as recommended in [8], it should not exceed about 5 s. It ensues that this type of listening is far from natural conditions for loudspeaker comparison.
- The second method is an alternate listening and enables the presentation of longer musical excerpts. The comparison is made by switching from one system to the other one; this is done by either an operator [3] or the listener himself [7]. This also enables one to test more than two loudspeakers in a single trial [9]. This approach is closer to the way music is listened to in real life.

2.3 Design of experiments

Taking these considerations into account, three paired-comparison experiments were designed. Successive and alternate presentations, described in 2.2, were proposed to the listeners. Even though the second presentation method appears as more natural for an average listener, the fact that he cannot control the reproduction level could be disturbing. In order to compare realistic comparison approach to laboratory listening tests, a third assessment procedure was designed. For that purpose, the listener was free to adjust the reproduction level before choosing his preferred loudspeaker.

3 Experiments

Three different assessment procedures were designed to gradually get close to real-life comparison approach. They all involved the same loudspeakers and music excerpts. Loudspeakers were presented in monophonic reproduction because it has proven to give more discriminating quality ratings than stereo or multichannel restitution [10]. The listeners had to take part to three sessions where the loudspeakers were presented by pairs. The same excerpt was played over the loudspeakers under test. The subject was asked to indicate his preferred loudspeaker.

3.1 Measurement scale

A preference rating scale had to be chosen to assess the answers. According to Toole [11], such a scale has to be used when the comparison of sounds relies on a relative basis rather than an absolute one. This kind of scale is considered by Jason [12] as intermediate in difficulty between a raw statement of preference and the IEC scale based on differences between fidelity judgements [2].

The continuous scale was divided into four equally wide intervals delimited with the labels indicated in Fig. 1 (translated from French).

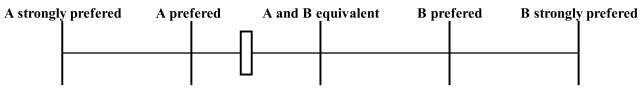


Fig.1 Answering scale of the assessment procedures.

The listener had to answer by moving a sliding cursor along the preference axis. The same scale was used during the whole test.

3.2 Stimuli

A stimulus denotes an excerpt reproduced over a loudspeaker. Four loudspeakers (namely A, B, C and D for the rest of the study) from different makers and assumed to be of about the same quality were then compared by pairs during the test.

A trial begins with the presentation of two stimuli and ends with their ratings. Then, if a single excerpt gives $N = 4$ stimuli, the number of trials for this excerpt is:

$$\frac{N(N-1)}{2} = 6 \qquad (1)$$

The result of a comparison of two loudspeakers is strongly dependent upon the content of the excerpts used as test material. The pieces of music chosen for loudspeaker comparison were:

- Leonard Bernstein, "West Side Story", symphonic orchestra.
- Ben Harper, "I want to be ready", human voice and acoustic guitar.
- George Gershwin, "Rhapsody in blue", piano solo.

They were extracted from CDs as 16-bit, 44.1-kHz wave format files. They were selected upon their ability to reveal spectral and preferential differences between different loudspeakers. The excerpts chosen for session involving an alternate presentation lasted 30 s. They were shortened to 5 s for consecutive presentation.

A session consisted of the 18 trials required to assess the 3 excerpts on the 4 loudspeakers: the first excerpt was proposed for the first 6 comparisons, the second excerpt was heard for the next 6 comparisons, and so on. It is worth noting that the three excerpts were presented in a random order over a session.

3.3 Test procedures

A test was divided into three different sessions described below:

- Session 1 was assumed to be the most repeatable and reliable; the 5-s stimuli were used. In one trial, the excerpt was successively presented on the loudspeakers, A and B. The subject was allowed to listen to the pair of stimuli as many times as needed before expressing his opinion through the measurement scale. Loudness was matched for this session.
- Session 2 was meant to be more consistent with everyday life listening. Longer stimuli (about 30 s) were proposed to the listener, who had the opportunity to switch, at any time, from loudspeaker A to loudspeaker B. He was allowed to listen to the excerpt and switch between both loudspeakers as many times as needed to make his opinion. The excerpt could be played from its beginning or from any other point chosen by the listener in the timeline, and the listening could be interrupted at any time. The question asked to listeners was always the same and was about their individual preference; loudness was also matched for this session.
- Session 3 was exactly the same as session 2, apart from the fact that loudness was not matched. The listener was allowed to vary the reproduction level of each loudspeaker under test by using a dedicated fader. The fader half-lift corresponded to the matched loudness, and the level could be varied from - 6 dB to + 6 dB around this value. At the beginning of each trial, the fader value was randomly set within these limits. The listener was

told to set the reproduction level at a comfortable value before making his comparison.

All of these 3 sessions were randomly presented to the listener over a test, which took about 1 h. Each session was preceded by a 5-min pre-test to familiarize the listener with the answering interface.

The instructions were given orally and in written form. Listeners were explicitly told to assess the stimuli according to their preference independently of their taste for the musical content.

3.4 Loudness matching

In this test, loudness was matched for sessions 1 and 2: a continuous pink noise was first used to roughly match the loudspeakers. The level was objectively equalized through adjustment of the gain control of each loudspeaker till getting 80 dB(B) at the listening position. Then, it was adjusted subjectively by three expert listeners so that the loudness sensation was the same for the 4 stimuli issued from a given excerpt. Each music excerpt was reproduced so that the level was close to the listening level preferred by the average listener [2].

3.5 Loudspeaker locations

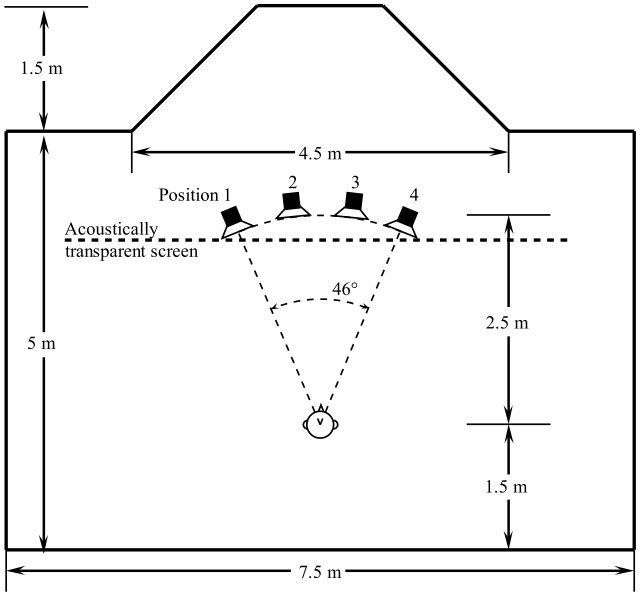
Since the purpose of this study was to get close to real-life conditions for loudspeaker comparison, the listeners could directly listen to sound radiations by loudspeakers rather than to recordings of these loudspeakers. Using direct presentation, two loudspeakers cannot be set exactly at the same position for comparison. The effects of loudspeaker positions can be higher than the subjective differences between the loudspeakers themselves [7]. On the other hand, Bech [13] noticed that, for most loudspeakers, the timbral quality of reproduced sounds is usually unaffected by changes in position within a radius of approximately 0.5 m.

Fig. 2 presents the listening room in use in this study. This room is a recording studio for amplified music. The listening position is at 1.5 m from the nearest wall. The same holds for all loudspeakers, which is in agreement with AES and IEC recommendations [1, 2]. All of the four loudspeakers are hidden behind a visually opaque, but acoustically transparent, screen. They are located at 2.5 m from the centre of the listener's head. The tweeters are placed at the height of the listener's ears. The distance between two contiguous tweeters is 0.5 m to keep interactions between them as low as possible while unaffecting the timbral quality.

As shown in Table 1, the reverberation time for this room, measured between 125 and 4000 Hz, ranges between 0.6 and 0.4 s, which agrees with IEC specifications [2].

f (Hz)	125	250	500	1000	2000	4000
RT (s)	0.6	0.58	0.55	0.52	0.45	0.4

Table 1 Reverberation time measured by octave bands in the listening room used for the tests.



.Fig.2 Listening room arrangement for monophonic loudspeaker comparison.

In agreement with IEC recommendations [2], loudspeaker positions were exchanged throughout the experiment to compensate for the positional influence over the preference ratings. The 4 loudspeakers were swapped between two listeners so as to test all of the 24 possible combinations of positions.

3.6 Listeners

The listeners involved in the test consisted of 5 women and 43 men aged from 20 to 60 years; the average age was 28. Most of them could be considered as trained listeners since they were sound engineering students or professionals working in the audio engineering field. Each of them carried out the three sessions consecutively.

Each loudspeaker layout was tested by two different listeners, so as to present twice all of the 24 possible combinations.

4 Results

4.1 Preference scores

The range of the preference scale was continuous and extended from -1 to +1 depending on whether the preference was marked for loudspeaker A or loudspeaker B, respectively. The answer to each trial was thus a preference probability within -1 and 1. Since listeners were free to use the answer scale at their convenience, no normalization was applied to the results. Let us denote P_{ij}, the preference probability of stimulus, i, versus stimulus, j. It is assumed that:

$$P_{ij} = -P_{ji} \tag{2}$$

A negative probability of preference P_{ij} means that the loudspeaker j is preferred to i. The linear relation described

by Eq. (3) allows one to derive the preference scores from these preference probabilities:

$$S_i = \sum_{j \neq i} P_{ij} \qquad (3)$$

where S_i is the preference score of stimulus i. As four loudspeakers were compared in this experiment, the preference score could theoretically lie within -3 and +3.

A multivariate analysis of variance was made to see whether the preference scores were affected by the loudspeaker, the loudspeaker position, the session and the excerpt.

4.2 Loudspeaker effect

The analysis of variance showed that the most influential factor was the loudspeaker itself ($F(3,1724) = 87.7$, $p < 0.0001$). However, the Fisher LSD test showed that, among the loudspeakers, only item A was significantly less appreciated than the three other ones ($p < 0.0001$). Fig. 3 shows that loudspeakers B, C and D were systematically preferred to loudspeaker A.

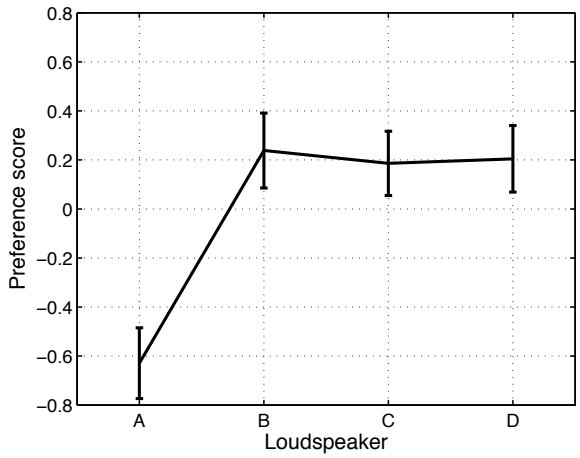


Fig.3 Mean preference scores for the four loudspeakers, within their 95% confidence interval.

4.3 Location effect

The loudspeaker location had also a significant influence on the preference scores ($F(3,1724) = 63.7$; $p < 0.0001$). Fig. 4 shows that, when the loudspeakers were set in front of the listener (positions 2 and 3 according to Fig. 2), their appraisal were significantly better than when they were on the sides (positions 1 and 4). This finding may result from different reasons: i) the listening room excitation depends on the loudspeaker location; ii) according to some listeners, the loudspeaker identification could be easier when the loudspeaker is set on the side (positions 1 and 4) and iii) when a loudspeaker is placed away from the listener's axis, he may need to turn his head towards the source, and this movement could have a negative effect upon his sound assessment.

No interaction was found between the loudspeaker and its location: the effect of the location was the same for the different loudspeakers.

Fig.4 Mean preference scores for the four positions, within their 95% confidence interval.

4.4 Loudspeaker/Session interaction

The interaction between the loudspeaker and the session was significant ($F(6,1721) = 3.06$; $p < 0.01$).

About the third session, Fig. 5 clearly shows the occurrence of more subtle preference ratings than the global scores displayed on Fig. 3. As previously observed for the loudspeaker effect, the only significant differences in sessions 1 and 2 were between item A and the three other ones ($p < 0.0001$). In session 3, the ratings about loudspeaker B were significantly different from the ones for loudspeakers C and D ($p < 0.05$ and $p < 0.001$, respectively). It means that the third session (where the listeners were allowed to adjust the reproduction level) was more discriminating than the two other ones.

It is worth noting that the preference scores for a given loudspeaker could significantly vary between two sessions. The scores obtained by Loudspeaker A during sessions 1 and 2 were significantly different ($p < 0.01$).

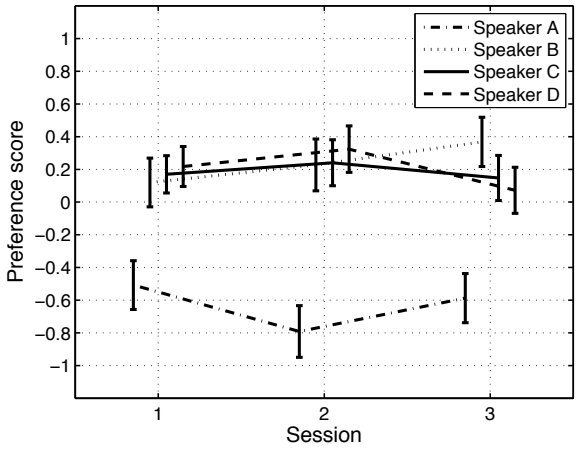


Fig.5 Mean preference scores for the four loudspeakers in the three sessions, within their 95% confidence interval.

4.5 Loudspeaker/Excerpt interaction

A significant interaction was also found between the loudspeaker and the excerpt ($F(6,1721) = 8.67$; $p < 0.0001$).

Loudspeaker A got the lowest preference scores, whatever the excerpt. But, loudspeaker C was significantly preferred to D ($p < 0.05$) for the 2nd excerpt (Ben Harper). For the third excerpt (Gershwin), loudspeakers B and D were significantly preferred to loudspeaker C (respectively $p < 0.001$ and $p < 0.0001$). This excerpt was a piano solo recording containing impulsive sounds. The piano was recorded live, and the recording was therefore a bit noisy. One should note that, with the Bernstein extract (symphonic orchestra with large masking effect), the loudspeakers B, C and D were equivalently rated.

5 Conclusion

This study dealt with the comparison of three different procedures designed to assess loudspeaker sound quality. The listeners were proposed successive or alternate listening, where the reproduction level was matched or left free. Successive and alternate presentation provided equivalent results when the loudness was matched. When given the possibility of setting the reproduction level, the listeners tended to give more discriminating assessments. Moreover the different tests showed significant effects of the extract and the loudspeaker position, which was already shown in previous studies.

Acknowledgments

The authors are grateful to the staff and students from *'Image & Son Brest'*, Université de Bretagne Occidentale.

References

[1] Audio Engineering Society, "AES20-1996: AES recommended practice for professional audio - subjective evaluation of loudspeakers (reaffirmed 2007)", *Journal of the Audio Engineering Society 44(5)*, 382-400 (1996)

[2] International Electrotechnical Commission, "Sound system equipment - Part 13: Listening tests on loudspeakers", *IEC Publication 60268-13* (1998)

[3] F.E. Toole, "Subjective measurements of loudspeaker sound quality and listener performance", *Journal of the Audio Engineering Society 33(1/2)*, 2-32 (1985)

[4] A. Gabrielsson, U. Rosenberg, H. Sjögren, "Judgments and dimension analyses of perceived sound quality of sound-reproducing systems", *Journal of the Acoustical Society of America 55(4)*, 854–861 (1974)

[5] E. Zwicker, H. Fastl, "Psychoacoustics – Facts and models", *Springer Verlag* (1990)

[6] B.C.J. Moore, B.R. Glasberg, T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness", *Journal of the Audio Engineering Society 45(4)*, 224-240 (1997)

[7] S.E. Olive, P.L. Schuck,, M.E. Sally, S.L. Bonneville, "The effects of loudspeaker placement on listener preference ratings", *Journal of the Audio Engineering Society 42(9)*, 651–669 (1994)

[8] M. Lavandier, P. Herzog, S. Meunier, "Comparative measurements of loudspeakers in a listening situation", *Journal of the Acoustical Society of America 123(1)*, 77–87 (2008)

[9] S.E. Olive, "A multiple regression model for predicting loudspeaker preference using objective measurements: Part 1 - Listening test results", in *Proceedings of the 116th AES Convention* (2004)

[10] F.E. Toole, "Subjective measurements of loudspeakers: A comparison of stereo and mono listening", in *Proceedings of the 74th AES Convention* (1983)

[11] F.E. Toole, "Audio engineering - Science in the service of art", in *Proceedings of the 111th AES Convention* (2001)

[12] M.R. Jason, "Design considerations for loudspeaker preference experiments", *Journal of the Audio Engineering Society 40(12)*, 979–996 (1992)

[13] S. Bech, "Perception of timbre of reproduced sound in small rooms: Influence of room and loudspeaker position", *Journal of the Audio Engineering Society 42(12)*, 999–1007 (1994)