A comparative study on different assessment procedures applied to loudspeaker sound quality

Vincent Koehl, Mathieu Paquier

A comparative study on different assessment procedures applied to loudspeaker sound quality

Vincent Koehl*, Mathieu Paquier

University of Brest, CNRS, Lab-STICC UMR 6285
6 avenue Victor Le Gorgeu, CS 93837
29238 Brest Cedex 3, France

Abstract

Paradoxically, one of the hardest to measure characteristics of a sound reproduction device such as a loudspeaker is its sound quality. The perception of this subjective character is linked to numerous parameters (stimulus type, listening environment...) that must be drastically controlled to lead to reliable and repeatable judgments. Industrial and academic researchers are still focusing on the design of standard assessment procedures. The conditions under which a sound reproduction system is assessed in laboratory tests is often very far from those under which it is designed to be used. As a result, the assessment task might appear unnatural to test subjects, which could possibly bias the test results. The aim of this study is to compare, on the basis of sound quality ratings, three different test procedures based on paired comparison and exhibiting procedural differences. One of the procedures consisted in comparing loudspeakers by listening to short music excerpts (5 s) at a preset level, which was assumed to be a very controllable method. In the two other procedures, the listener could compare the systems by listening to long music excerpts (30 s), which was assumed to be more natural for loudspeaker assessment. The level was either preset by expert listeners or set by the subject himself in the two latter procedures. This paper shows that the test results were very stable over the different assessment procedures, but that some of them enabled, under certain conditions, to separate between systems obtaining very close quality ratings.

*Corresponding author

Email addresses: vincent.koehl@univ-brest.fr (Vincent Koehl), mathieu.paquier@univ-brest.fr (Mathieu Paquier)

Keywords: Loudspeaker sound quality, Assessment procedure, Paired comparison

1. Introduction

The way a listener determines the sound quality of a product is a very complex phenomenon. According to Jekosch [1], the perceived quality results from a comparison process: the subject assesses what is actually presented to him with respect to a "desired" stimulus. As an example, the best sound quality rating is reached when the system under test fulfil the listener's expectations towards this item. Moreover, for the specific case of sound reproduction devices, the result of the quality assessment relies not only on the system under test itself but also on numerous external factors such as the stimulus type and the listening environment [2, 3, 4]. As these parameters may interact with the system under test, they could modify the way it is perceived and affect the sound quality. As a result, the subjective characterization of loudspeakers is a difficult and time-consuming task. The parameters to be controlled are numerous, see Toole [5] for a review, and the results of such experiments are often very context-dependent. The Audio Engineering Society and the International Electrotechnical Commission have both provided recommendations dedicated to the subjective evaluation of loudspeakers [6, 7]. The International Telecommunication Union has also published a wide range of recommendations for the subjective evaluation of sound quality. The main guidelines of these recommendations are summarized in ITU-R BS.1284 [8].

Finally, the controlled experimental conditions needed to obtain reliable and repeatable quality judgments from a listening test are very far from the conditions under which a loudspeaker is intended to be used. As an example, assessing loudspeaker sound quality by listening to short stimuli at a predefined level might appear rather unnatural to subjects. In this section, the main issues of loudspeaker sound quality evaluation are reviewed and discussed in order to design assessment procedures and to compare them on the basis of the sound quality ratings they might provide about different loudspeakers.

1.1. Stimulus presentation

It is often recommended to present the systems under test (loudspeakers or whatever) in a multiple presentation layout (at least by pairs) to facilitate the evaluation. Even when absolute judgments are wanted, the loudspeakers are often presented by pairs to ease the task. This procedure is referred to as paired ratings by IEC [7]. According to Toole [9], comparisons must be as quick as possible to ensure a maximum discrimination and a minimum variability in the assessments. Paired comparison appears then as a reliable way to estimate loudspeaker sound quality. It can be accomplished either by listening consecutively to the two systems under test or by switching from one system to the other one at any time. This first presentation method is often referred to as *AB* comparison, the stimuli have to be heard one just after the other [2]. This procedure appears as the most exact way to compare two stimuli and is therefore often recommended for psychoacoustic experiments [10]. In the second presentation method, the switching process can be done by either an operator [9] or the listener himself [2]. This kind of procedure makes also possible the assessment of more than two systems in a single trial using multiple comparisons [11] and is then referred to as MUSHRA (MUlti Stimulus test with Hidden Reference an Anchor) by ITU [12] and is often called so even when used without reference or anchor [13, 14].

1.2. Stimulus type

According to previous studies [2, 3, 4], the subjective assessment is strongly dependent on the content of the excerpts used as test material. Therefore, it is advised to perform the test on various excerpts that exhibit different spectral and temporal features. Short stimuli should be preferred for *AB* comparison because of the human auditory memory [2]. As recommended by Lavandier et al. [15], their duration should not exceed about 5 s. To compare loudspeakers using longer stimuli, *AB* comparison should be avoided because of the difficulty of memorizing long sequences. An alternate listening with switching possibility shall be preferred. In any case, this approach, based on long stimuli, is more consistent with a natural experience of music listening. However, when enabling the listener to switch between the systems at any time, one cannot exclude that he will focus his attention on a specific part of the sequence which might be in addition different from one subject to another.

1.3. Loudness matching

To compensate for the loudness influence over the subjective assessments [16], loudness-matched stimuli are usually presented in listening tests. A common prerequisite to the subjective evaluation is then to check that the perceived reproduction level (i.e. loudness) is alike for all of the loudspeakers under test. The two well-known loudness models developed by Zwicker and Fastl [10] and Moore et al. [17] were designed for steady-state sounds. Modified versions of these models have been developed over the past years [18, 19] for time-varying signals and proved their efficiency on synthetic and technical sounds [20]. However, such model-based estimates are still rarely considered for long music excerpts that are commonly used to assess loudspeaker sound quality as they may exhibit large loudness fluctuations throughout their durations.

As a matter of fact, the loudness matching is often finally accomplished subjectively by the experimenter himself or by several expert subjects prior to the listening tests [6]. The level is then set to a value that is assumed to be the preferred listening level for the average listener [7], according to the type of stimulus (generally music) under test, and matched over the different systems under test. However, the fact that the loudspeakers are presented at a preset level might appear unnatural to the listener who could possibly wish to choose it himself.

1.4. Listening environment

When several loudspeakers are presented to listeners for direct comparison, they cannot be set exactly at the same position in the listening room. This fact is a matter of concern: the effects of loudspeaker positions over subjective assessments can be higher than the intrinsic differences between the loudspeakers [2]. Although Bech [4] 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, the positioning issue in listening tests about loudspeakers is still subject to debate. To tackle this problem, experimenters often set and record the different systems at a given position by using generally binaural techniques [21], the recordings being presented to the listeners over headphones. Recent studies showed that loudspeaker comparison using binaural recordings proved to be an efficient alternative to direct comparison in numerous cases [22, 23]. Another way of avoiding the effects of the loudspeaker position over the quality judgment is the use of a

setup that enables to rapidly switch the loudspeakers over the same position like a turntable [9] or an automatic shuffler [24]. With such setups, the transition time is hardly below 1 s (2 s in average for the automatic shuffler), which is already too much if the listener should be able to instantaneously switch between systems to be compared.

1.5. Judgment scaling

Loudspeakers are often assessed in terms of "fidelity" [7]. However, this notion may cause a confusion in the listener's mind. The typical stimuli for the perceptual assessment of loudspeaker are music excerpts for which the listener might have no clue about how they should originally sound. The subject might wonder whether the fidelity should be related to the original excerpt or to the original sound scene, on which he has expectations but few objective information. The more so that this scene might be strongly modified by numerous parameters such as the recording room acoustics or the mixing process. The loudspeaker itself may modify the original input because of its own frequency response [25]. Gabrielsson et al. [26] have shown that a slight amplification of the medium to high frequencies can actually lead to increased fidelity ratings compared to the "flat" response condition. This finding suggests that the listeners' desired reproduction (internal reference) did not match the original one in this case. As a result, the colorations added by the resonances occurring even within a high-quality loudspeaker might be perceived [27] and, depending on the test stimulus, enhance or degrade its perceived fidelity. As a result, the fidelity is assessed according to individual taste and experience. The best fidelity then is reached when the reproduced stimulus matches the listener's expectations, as for perceived sound quality. Therefore, assessing a loudspeaker about its fidelity is often considered as equivalent to evaluating its sound quality [28].

In multiple stimulus presentation, the listeners are asked to rate the systems under test along either a fidelity scale [7] or a quality one [12]. For this kind of assessment, the listener has to grant each system under test an absolute mark, in contrast to preference judgments where he only has to state or quantify his preference. Zielinski et al. [29] recently addressed the numerous biases related to the use of quality or fidelity scales and recommended indirect scaling method through, for example, paired comparisons. The subjective assessment can then be carried out on a preference scale and focus only on the relative performance of the devices. According to Jason [30],

a judgment on a preference scale is considered as intermediate in difficulty between a raw statement of preference and the IEC fidelity scale, from which a preference judgment can be derived, if needed, as the difference between two fidelity judgments. Although paired comparisons on a preference scale are time consuming, this procedure does not require any explanation on the concepts of quality or fidelity on which the listeners do not necessarily agree.

1.6. Summary

The aim of this paper is to compare assessment tasks exhibiting procedural difference on the basis of the sound quality ratings they provide about different loudspeaker models. All the considerations developed above were taken into account to design three different test procedures, all based on paired comparisons. The loudspeakers were thus assessed on a relative basis, which means that the listeners had to indicate their preference over the pair under test and not to grant each system a quality or fidelity mark. The sound quality ratings for each loudspeaker under test were derived from these preferences. The comparisons were achieved on different kinds of music excerpts, as the result of a test is acknowledged to be highly correlated to the program material used for the evaluation. The physical performances of the loudspeakers will not be investigated since the aim of the study was not to link the objective performances to the subjective ones.

The *AB* and switched comparison procedures described above were investigated, enabling to present respectively short and long music excerpts to the listeners. In a first procedure, the listener's task was to compare two short stimuli (matched in loudness) by listening to them consecutively (*AB* comparison). This method will be referred to as consecutive presentation for the rest of the paper. Although this listening situation is far from the way loudspeakers are naturally used, it ensures repeatable comparisons. It appeared worthwhile comparing sound quality ratings provided by such a procedure to those obtained from assumed more natural procedures, where the comparison can be achieved by listening to longer music excerpts at a level that is not necessarily set beforehand. Therefore, two other procedures were tested, for which the subject had to compare two long stimuli by listening to them alternatively (switching was allowed at any time). This method will be referred to as alternate presentation for the rest of the paper. The loudness was matched in one of these two test procedures, in the other one

the listener was free to adjust the reproduction level for each of the two systems under test before giving his preference.

2. Experimental setup

The perceived audio quality of four loudspeakers was measured in a listening test including three different presentation methods. These procedures were various trade-offs between the reliability and the "naturalness" of the comparison task. The loudspeakers under test came from different manufacturers and were presented to the listeners in monophonic reproduction because it is acknowledged to be more discriminating than stereophonic or multichannel reproduction for sound quality evaluation [31, 32]. Assessment procedures that would appear as natural to the listeners were under study and a comparison of loudspeakers through headphones was thus not considered. It was decided that the subjects had to listen to direct sound radiations rather than to recordings of the loudspeakers. The systems under test were from the same price range and exhibited clearly audible timbre (and presumably quality) differences. They were tested without any a priori knowledge on the quality ratings that they might obtain using the assessment procedures under test. The four tested loudspeakers are denoted by Ls_1, Ls_2, Ls_3 and Ls_4 for the rest of the paper.

2.1. Scaling method

An indirect scaling method was chosen to evaluate the loudspeaker sound quality. The listener was asked about his preference between two systems instead of absolute quality judgments. Here, the subject could indicate his preference over the two loudspeakers involved in a paired comparison along a single continuous scale on a MATLAB graphical user interface. The two current loudspeakers were always respectively denoted by A and B on the screen, whatever the presentation method (consecutive or alternate). The scale was divided into four intervals of equal width delimited by the labels indicated in Fig. 1 (translated from French).

2.2. Program material

In order to compare two different loudspeakers, the same excerpt was played (consecutively or alternatively depending on the procedure) over them.

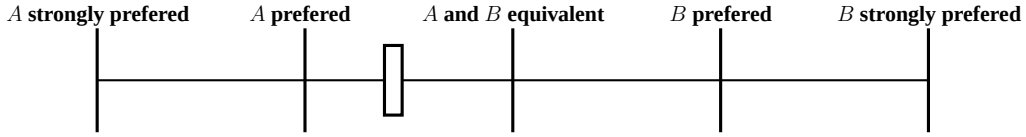


Figure 1: Preference scale in use during the paired comparisons.

Table 1: List of musical program material.

Excerpt	Disc	Track	Time (long excerpt)	Time (short excerpt)
Ex_1	Los Angeles Philharmonic Orchestra: -Rhapsody in blue (George Gershwin) -West side story (Leonard Bernstein)	6. Mambo (Presto)	0'10"-0'36"	0'19"-0'23"
Ex_2	Ben Harper: -The will to live	6. I want to be ready	0'07"-0'41"	0'07"-0'12"
Ex_3	Los Angeles Philharmonic Orchestra: -Rhapsody in blue (George Gershwin) -West side story (Leonard Bernstein)	1. Rhapsody in blue	8'48"-9'18"	8'48"-8'53"

As indicated in Table 1, three music excerpts were selected from commercially available stereo material. According to their different musical contents (symphonic orchestra, human voice and acoustic guitar, piano solo), they were selected upon their ability to reveal preferential differences between different loudspeakers.

For each piece of music, a rather long excerpt (around 30 s) was initially selected. Then a short excerpt (around 5 s) was extracted from each of these selections. The long and the short excerpts were respectively selected for the alternate presentation and the consecutive one. Each short excerpt was selected upon agreement between three expert listeners (the two authors and one loudspeaker designer). It was assumed to be as representative as possible of its long version (i.e. perceived as having equivalent spectral and dynamic features) to reveal the same perceptual differences between the loudspeakers. All excerpts were extracted from CDs as 16-bit, 44.1-kHz wave-format files.

These stereo excerpts were mixed down to mono by adding the left and right channels. It is often advised to use only one channel for monophonic evaluations [6], because of the possible cancellations of certain signal components. Nevertheless, it appeared to us that listening to only one side of the stereo scene would not sound natural to the subject, especially with the symphonic orchestra. Comparative listening (between the stereophonic

reproduction of the initial signals and the monophonic reproduction of the reduced signal) showed no corruption of the excerpts by this addition. Similar stereo-to-mono reduction was achieved by Choisel and Wickelmaier [33], as recommended by ITU-R BS.775-1 [34], in order to compare the stereophonic and the monophonic reproductions of multichannel recordings.

One should note that, here, a stimulus denotes an excerpt reproduced over a given loudspeaker. Presentation for each excerpt then results in $N = 4$ stimuli, the stimuli numbers being randomized independently for each excerpt, listener and assessment procedure. A trial began with the presentation (consecutive or alternate) of two stimuli and ended with the preference rating. The number of trials needed to achieve all possible paired comparisons of this 4 stimuli was (without comparing a stimulus to itself):

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

These 6 pairs were arranged according to a Ross' series [35] to avoid as much as possible the successive presentation of two pairs having one stimulus in common. The 6 paired comparisons involving one excerpt were presented in a row to the listener to facilitate his task. A session consisted then of the 18 trials required to assess the 3 excerpts: in practice, the first excerpt was submitted to the first 6 comparisons, then, the second one was heard for the next 6 comparisons and finally the last one was presented for the last 6 comparisons. The excerpt order was randomized for each session.

2.3. Loudness matching

For each music excerpt indicated in Table 1, an assumed realistic listening level was chosen upon agreement between the three expert listeners that selected the stimuli.

The 4 loudspeakers under test were then matched so that the loudness related to the 4 stimuli issued by a given excerpt was the same. This matching process was firstly objectively accomplished using a continuous pink noise and secondly subjectively confirmed [6]. At the first stage, the level was equalized through adjustment of the gain control of each loudspeaker till obtaining 80 dB (B) at the listening position. This matching then was subjectively confirmed by each of the three expert listeners by listening to the

Table 2: Main features of the 3 procedures of the listening test.

Procedure	Presentation method	Excerpt type	Reproduction level
Pr_1	-consecutive	-short	-matched in loudness
Pr_2	-alternate	-long	-matched in loudness
Pr_3	-alternate	-long	-set by the listener

music excerpts at their respective preselected listening levels.

As an indication of these listening levels, level measurements (averaged over the duration of the stimuli) at the listening position were around 80.5 dBSPL for excerpt 1, 76.2 dBSPL for excerpt 2 and 73.4 dBSPL for excerpt 3. These values could slightly vary from one loudspeaker to another, and also according to the excerpt duration, but these variations were generally less than 1 dB.

2.4. Assessment procedures

For each subject, the test was made of three sessions, each corresponding to one of the three assessment procedures described below and denoted, respectively, by Pr_1, Pr_2 and Pr_3. Their main features are summarized in Table 2. It is worth recalling that the three excerpts were tested in each session (i.e. 18 trials per session). The procedure order was randomized over tests and each session was preceded by a 3-min pre-test to familiarize the listener with the answering interface and the stimuli. The question asked to listeners was the same in each session and was about their individual preference within pairs; instructions were given orally and in written form. The listeners were explicitly told to assess the stimuli according to their preference independently of their taste for the musical content. During a trial, the currently auditioned stimulus was always indicated on the screen. As recommended by ITU [12], while the subject was listening to a stimulus, all of the on-screen objects that corresponded to the other stimulus were disabled to prevent mistakes. A 2-min break was given to the listener between two sessions and the test took altogether about 1 h.

2.4.1. Procedure Pr_1

This procedure was assumed to be the most repeatable and reliable one; short stimuli (matched in loudness beforehand) were used for consecutive presentation. In one trial, the excerpt was consecutively played over the two

loudspeakers to be compared (AB comparison). The subject was allowed to listen to the pair of stimuli as many times as needed before reporting his opinion on the measurement scale. For a given subject, each stimulus pair was only presented in one given order (and not in the reverse one) to shorten the session and to obtain the same number of trials as the two other procedures (as explained before the 4 stimuli were first randomized and then the 6 pairs were specifically arranged). It was assumed that the stimulus presentation order would have no significant effect on the preference ratings. Moreover, the possible order effect was balanced over the whole listener panel because of the random arrangement of the stimuli.

2.4.2. Procedure Pr_2

This presentation was meant to be more consistent with the conditions under which a loudspeaker is used: long stimuli (matched in loudness beforehand) were proposed to the listener in alternate presentation. The subject had the opportunity to switch, at any time, from one loudspeaker to the other. 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 that was available on the screen. The listening could be interrupted at any time.

2.4.3. Procedure Pr_3

In order to place the listener in a natural situation to assess of loudspeaker sound quality, it was decided to finally let him set his own preferred listening level. This procedure was then identical to the previous one, 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 two faders were displayed on the screen. The reproduction level corresponding to the matched loudness was assigned to the center of the fader stroke and the level could be varied from -6 to $+6$ dB around this value. The listener was told to set the reproduction level at a comfortable value before giving his judgment. In order to encourage him to vary the volume, the two faders were randomly set within their limits at the beginning of each trial. All level settings were stored for further analysis.

2.5. Listening room

The listening room was a recording studio for amplified music. Table 3 indicates that its reverberation time, measured between 125 and 4000 Hz,

lies within 0.6 and 0.4 s. This reverberation time complies with the IEC 60268-13 standard specifications [7] (Fig. 2).

Table 3: Reverberation time measured by octave bands in the listening room used for the tests.

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

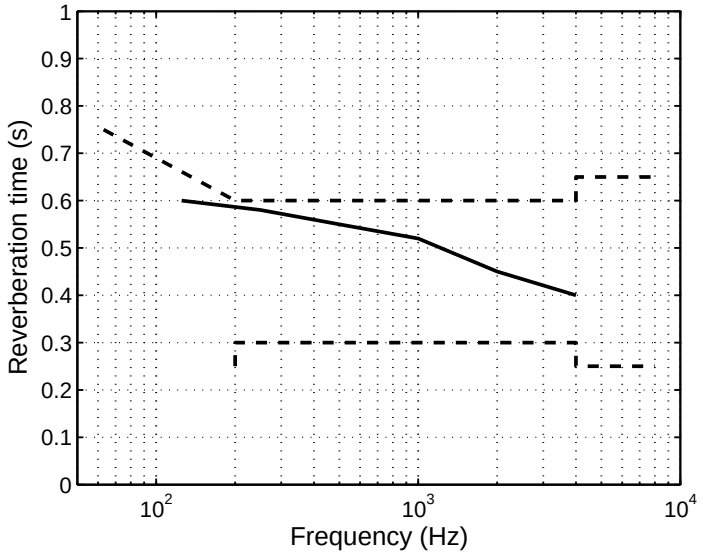


Figure 2: Reverberation time for the listening room within the IEC 60268-13 tolerances.

Fig. 3 shows that the subject and the loudspeakers were all set at 1.5 m from the nearest wall, in agreement with AES and IEC recommendations [6, 7]. The 4 loudspeakers were hidden behind a visually opaque, but acoustically transparent, screen. They were located at 2.5 m from the center of the listener's head which was not constrained. The tweeters were placed at the height of the listener's ears. The distance between two contiguous tweeters was 0.5 m to keep interactions between the loudspeakers as low as possible. In this case, two stimuli to be compared could then be generated by 1.5-m-distant loudspeakers. The listener had thus to compare loudspeakers for which the positional effect might not be negligible [4] but could instantaneously switch between them. The loudspeaker positions were fixed for a

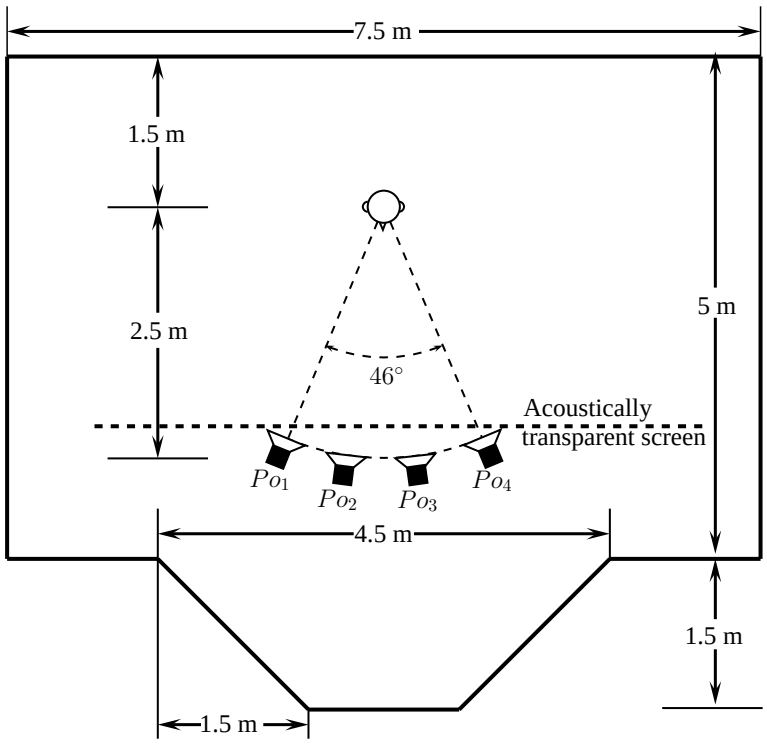


Figure 3: Listening room arrangement for monophonic loudspeaker comparison.

given listener and then exchanged from one subject to another to compensate for this effect over the whole listener panel. The 4 loudspeakers were thus swapped across the 4 positions (Po_1 to Po_4 according to Fig. 2) so as to assess all of the $4! = 24$ possible combinations throughout the study.

2.6. Listeners

Fourty-eight listeners (5 women and 43 men) aged from 20 to 60 years (average 28) participated in this experiment. Most of the subjects (37 out of 48) were sound engineering students (Master's degree) from the University of Brest who had passed an audiogram in the month preceding the test and showed normal hearing thresholds. The sample group was complemented by professionals working in the audio engineering field who reported normal hearing. None of them had particular experience in laboratory listening tests. However, because of their work experience, personal knowledge and interest, they were considered as competent to evaluate the sound quality of

loudspeakers. According to ISO 8586-2 standard [36], they match the definition of "expert". Each subject carried out the three assessment procedures consecutively but in random order, with the four loudspeakers arranged according to a given layout that was fixed for his test.

In all, each of 24 different loudspeaker arrangements was tested by two different listeners, so as to assess twice each specific combination. Listeners' effects over the quality ratings were thus not regarded; because of their expert status, they were expected to exhibit low variability in their subjective assessments. The influence of the position over the preference ratings could be investigated instead.

2.7. Derivation of the quality ratings

The preference scale (Fig.1) was continuous and numerically ranged from 0 to 1. Since listeners were free to use the answering scale at their convenience, no normalization was applied to the results. The preference of stimulus i versus stimulus j is denoted by P_{ij} where $P_{ij} = 1$ indicates in this case a strong preference for stimulus i and $P_{ij} = 0$ a strong preference for stimulus j. P_{ji} can be deduced by using:

$$P_{ji} = 1 - P_{ij} \tag{2}$$

The result to each trial was thus a preference rating lying within 0 and 1. For each subject, preference ratings (related to stimulus pairs) were then transformed into sound quality ratings (related to single stimuli) by using a linear model to compute merit scores [37]:

$$S_i = \sum_{j \neq i} P_{ij} \tag{3}$$

where S_i is the merit score of stimulus i. The merit scores related to the 4 stimuli needed to reproduce a given excerpt over the different loudspeakers were obtained by using the 6 preferences ratings needed to achieve the paired comparisons involving these stimuli, as explained in Eq. (1). Using this method each merit score, denoted as sound quality rating for the rest of the paper, lies thus within 0 and +3. The linear model enabled a good correlation ($r = 0.85$; $p < 0.001$) between measured preferences and the ones that were reconstructed by using the scores obtained from Eq. (3).

3. Results

The effective test duration (without pre-tests and breaks) was in average 44'43", where the mean durations per session were:

- 08'42" for procedure 1,
- 16'33" for procedure 2,
- 19'28" for procedure 3.

A four-way analysis of variance (Table 4) was carried out to look at the effects of the factors:

- Pr: assessment procedure (3 levels),
- Ls: loudspeaker model (4 levels),
- Po: loudspeaker position (4 levels),
- Ex: music excerpt (3 levels),

and their interactions over the sound quality ratings. One should note that simple effects of the procedure and the excerpt were not directly observable, the average rating per procedure and per excerpt being always equal to 1.5 because of the data collection method (i.e. ratings derived from preferences). As an example, considering only one excerpt and one listener, the average score $\bar{S}$ in each procedure is necessarily:

$$\begin{aligned} \bar{S} &= \frac{S_1 + S_2 + S_3 + S_4}{4} \\ &= \frac{(P_{12} + P_{13} + P_{14}) + (P_{21} + P_{23} + P_{24}) + (P_{31} + P_{32} + P_{34}) + (P_{41} + P_{42} + P_{43})}{4} \\ &= \frac{6}{4} \end{aligned} \tag{4}$$

and shall remain the same when averaging over excerpts and listeners. Simple procedure and excerpt effects could have been observed by collecting absolute judgments but their significance would only imply that the stimuli are globally more appreciated using one procedure or excerpt than another one. The present experiment is neither designed nor aimed at highlighting such effects. It is here worth recalling that the goal of the present study is to compare listening procedures on the basis of the sound quality evaluations

Table 4: Results of the four-way analysis of variance on sound quality ratings.

Source	SS	DF	MS	F	p[a]
Pr	0	2	0	0	1
Ls	57.121	3	19.0404	87.70	0***
Po	41.483	3	13.8277	63.69	0***
Ex	0	2	0	0	1
$Pr \times Ls$	3.983	6	0.6638	3.06	0.0056**
$Pr \times Po$	0.328	6	0.0547	0.25	0.9587
$Pr \times Ex$	0	4	0	0	1
$Ls \times Po$	3.626	9	0.4029	1.86	0.0544
$Ls \times Ex$	11.290	6	1.8817	8.67	0***
$Po \times Ex$	5.660	6	0.9435	4.35	0.0002***
$Pr \times Ls \times Po$	2.120	18	0.1178	0.54	0.9388
$Pr \times Ls \times Ex$	5.833	12	0.4861	2.24	0.0085**
$Pr \times Po \times Ex$	2.194	12	0.1828	0.84	0.6068
$Ls \times Po \times Ex$	6.084	18	0.3380	1.56	0.0633
$Pr \times Ls \times Po \times Ex$	3.615	36	0.1004	0.46	0.9974
$Error$	343.900	1584	0.2171		
$Total$	487.238	1727			

they provide about four different loudspeakers. Such a comparison can still be achieved by looking at the interaction between procedure and loudspeaker.

3.1. Loudspeaker effects

According to the analysis of variance, the most influential factor is the loudspeaker itself ($F(3,\ 1584) = 87.7$; $p < 0.001$). The loudspeakers under test obtained thus statistically different sound quality ratings. However, according to Fisher's LSD test[b], among the loudspeakers, only item Ls_1 was significantly ($p < 0.001$) less appreciated than the 3 other ones; Fig. 4 evidences that the loudspeakers Ls_2, Ls_3 and Ls_4 were systematically preferred to it.

[a] Where * stands for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$, as for the rest of the paper.

[b] Fisher's LSD test was used for all post-hoc analyzes in the present paper.

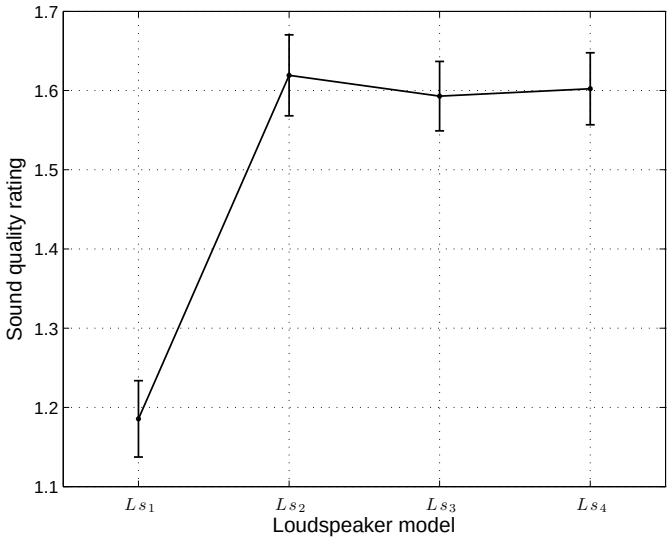


Figure 4: Mean quality ratings for the 4 loudspeakers, within their 95% confidence intervals.

The sound quality rating related to a given loudspeaker model proved to be dependent on the way it was assessed, as shown by the significant $Pr \times Ls$ interaction ($F(6,\ 1584) = 3.06$; $p < 0.01$). Using procedure Pr_3, Ls_2 could be statistically separated from Ls_3 and Ls_4, whereas these 3 loudspeakers could not be distinguished using Pr_1 and Pr_2. As shown in Fig. 5, Ls_2 obtained significantly higher ratings than Ls_3 and Ls_4 ($p < 0.05$ and $p < 0.01$ respectively) in the third procedure.

The sound quality rating related to a given loudspeaker model was also dependent on the excerpt used for the comparison, as shown by the significant $Ls \times Ex$ interaction ($F(6,\ 1584) = 8.67$; $p < 0.001$). Whatever the excerpt, Ls_1 obtained significantly lower ratings than its three opponents. Excerpts Ex_2 and Ex_3 proved to be more discriminant than Ex_1 (Bernstein, symphonic orchestra with possibly large masking effect) using which Ls_2, Ls_3 and Ls_4 could not be statistically separated. Using Ex_2 (Ben Harper), loudspeaker Ls_3 obtained a significantly higher rating than Ls_2 ($p < 0.05$) and Ls_4 ($p < 0.01$), as shown in Fig. 6(a). On the contrary, Ls_3 obtained significantly lower quality ratings than Ls_2 and Ls_4 ($p < 0.001$ in both cases) using Ex_3 (Gershwin, piano solo recording with impulsive sounds), see Fig. 6(b).

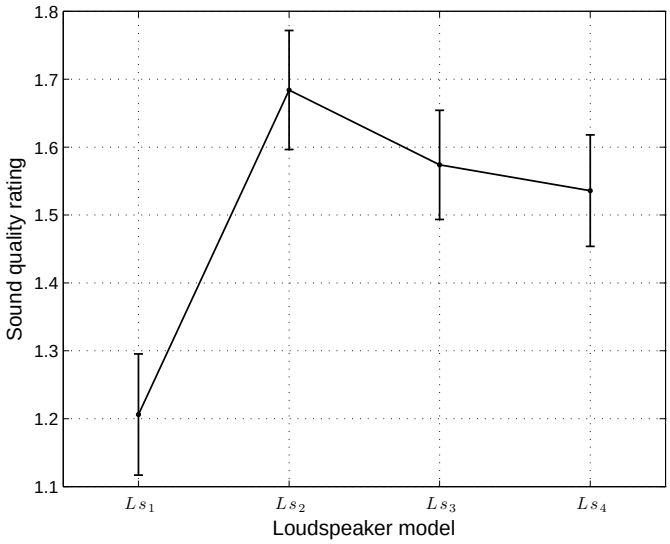


Figure 5: Mean quality ratings for the 4 loudspeakers obtained by using assessment procedure Pr_3, within their 95% confidence intervals.

The fact that a loudspeaker quality rating may vary across the excerpts indicates that the loudspeaker behavior (and subsequently its assessment) depends on its excitation, which was already shown by past studies [2, 3, 4].

Finally, the sound quality rating for a given loudspeaker was dependent on the combination of the music excerpt and the assessment procedure, as proven by the significance of the $Pr \times Ls \times Ex$ third-order interaction ($F(12,\ 1584) = 2.24$; $p < 0.01$). As an example, the observations that were made about Ex_2 and Ex_3 when considering all procedures (see respectively Fig. 6(a) and Fig. 6(b)) were also valid when considering only Pr_2 as illustrated in Fig. 7(a) and Fig. 7(b) respectively. This applied also when considering only Pr_3 but not for Pr_1 for which only Ls_1 could be significantly separated from the three other loudspeakers. In addition, whatever the excerpt-procedure combination, loudspeaker Ls_1 obtained significantly lower ratings than the other items under test.

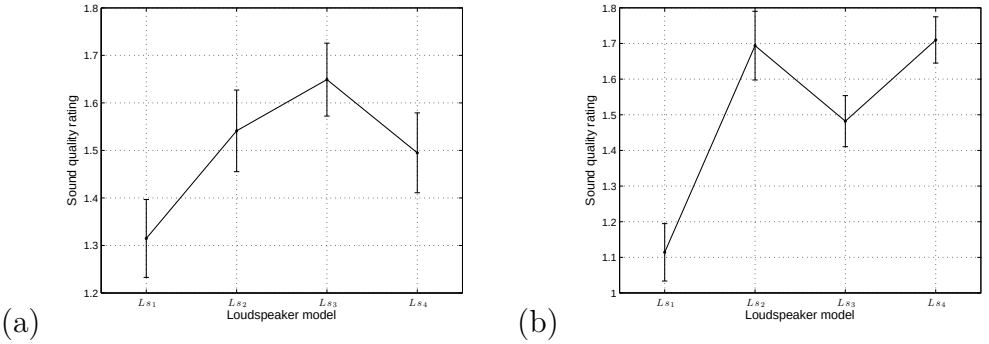


Figure 6: Mean quality ratings for the 4 loudspeakers obtained by using excerpts Ex_2 (a) and Ex_3 (b), within their 95% confidence intervals.

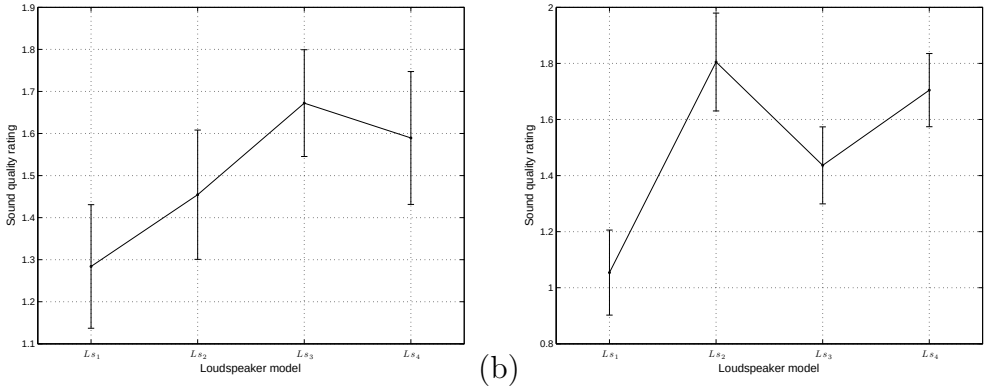
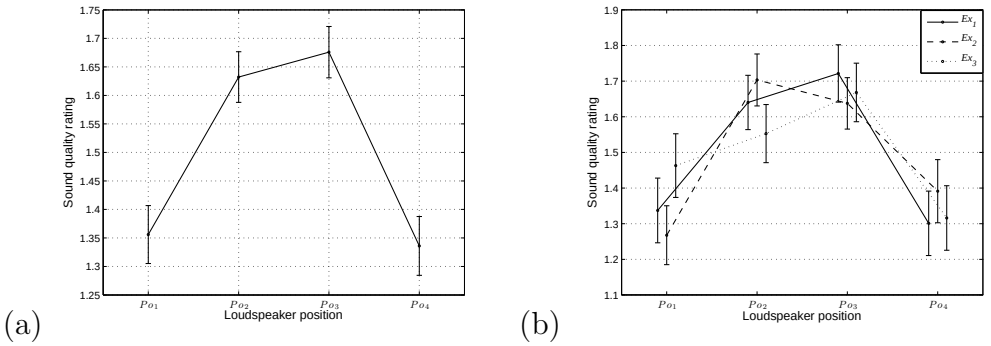


Figure 7: Mean quality ratings for the 4 loudspeakers obtained by using assessment procedure Pr_2 and excerpts Ex_2 (a) and Ex_3 (b), within their 95% confidence intervals.

3.1.1. Position effects

As expected from previous studies [2, 4], the loudspeaker position also had a significant influence on the perceived sound quality ($F(3,\ 1584) = 63.69$; $p < 0.001$). When the loudspeakers were placed in front of the listener (Po_2 and Po_3 with respect to Fig. 3), the sound quality ratings were significantly higher than when placed on the sides (Po_1 and Po_4), as can be seen in Fig. 8(a).

The loudspeakers were thus preferred in frontal position. This statement may have several explanations: (i) the listening room excitation depends on the loudspeaker location and (ii) when the loudspeaker is placed off-axis, the



(a) (b)

Figure 8: Mean quality ratings for the 4 positions (a) and for each excerpt (b), within their 95% confidence intervals.

listener may need to turn his head toward the source and this "effort" could have a negative effect on his quality assessment. No significant interaction was found between the loudspeaker and the position: the positional effect was the same for the different loudspeakers.

The positional effect was dependent on the excerpt as shown by the significance of the interaction between these two parameters ($F(6,\ 1584) = 4.35$; $p < 0.01$). In a given position, the quality ratings obtained using different excerpts might be significantly different, as illustrated in Fig. 8(b). This suggests that the listening room response depends on the loudspeaker-excerpt combination. According to Bech [4], the differences between positions could be more audible for certain programs.

4. Discussion

The main observation that can be made about this analysis is that the sound quality ratings regarding the four loudspeakers under test were very stable over the different assessment procedures and music excerpts. As a rule, loudspeaker Ls_1 obtained systematically significantly lower quality ratings than Ls_2, Ls_3 and Ls_4 which were practically judged equivalent. The three latter loudspeakers could be separated only under certain conditions: using procedures Pr_3 whatever the excerpt and using Pr_2 with excerpts Ex_2 and Ex_3.

It could then be thought than the long excerpts used in procedures Pr_2

and Pr_3 were more suited to the evaluation of loudspeaker sound quality than the short ones as they enabled slightly but significantly finer assessments. Based on the three music excerpts under test, such durations should then be preferred when loudspeaker of very close sound qualities have to be compared. However, some differences can be noted in the assessments stemming from procedures Pr_2 and Pr_3 involving these long excerpts. Pr_3 appeared slightly more discriminant as it enabled to separate between loudspeakers having similar sound qualities when taking all excerpts into account, whereas Pr_2 enabled it only for 2 excerpts among 3. The only procedural difference was that the listener was free to adjust the relative reproduction level (denoted by L), at his convenience between -6 and $+6$ dB around the value that corresponded to the initially matched loudness ($L = 0$ dB), throughout Pr_3 as indicated in 2.4.3. For each trial, the listener could adjust L for the two stimuli involved in the paired comparison by using dedicated faders displayed on the screen. He was told to indicate his preference once the two levels were set. Nevertheless, it is worth wondering what caused the differences between the two alternate presentations: is it caused by the supposed naturalness of Pr_3 or by an artifact of loudness differences, introduced by the test procedure, and leading to differences in the preference ratings? According to Gabrielsson et al. [16], when two equivalent loudspeakers are compared at different levels, the loudest one can be preferred.

A three-way analysis of variance was then carried out to examine how the relative reproduction level L was affected by the 3 experimental variables:

- Ls: loudspeaker model (4 levels),
- Po: loudspeaker position (4 levels),
- Ex: music excerpt (3 levels),

and by possible interactions in procedure Pr_3.

This analysis (see Table 5) showed that the factor loudspeaker had no significant influence on the level setting; the average listening level was alike for the 4 different loudspeakers under test. The preferred level for a given loudspeaker did also not depend on the excerpt used to achieve comparisons, as proven by the non-significant $Ls \times Ex$ interaction. On the other hand, the excerpt effect proved to be highly significant ($F(2,\ 1680) = 13.30$; $p < 0.001$);

Table 5: Results of the three-way analysis of variance on level settings.

Source	SS	DF	MS	F	p
Ls	20.3	3	6.771	0.79	0.5018
Po	13.5	3	4.516	0.52	0.6657
Ex	229.2	2	114.620	13.30	0***
$Ls \times Po$	399.4	9	44.377	5.15	0***
$Ls \times Ex$	107.2	6	17.874	2.07	0.0533
$Po \times Ex$	19	6	3.166	0.37	0.8998
$Ls \times Po \times Ex$	237.3	18	13.185	1.53	0.0709
$Error$	14475.4	1680	8.616		
$Total$	15501.5	1727			

the differences in level settings among excerpts can be seen in Fig. 9.

The listening level was also dependent on the loudspeaker-position combination, as shown by the significant $Ls \times Po$ interaction ($F(9,\ 1680) = 5.15$; $p < 0.001$). At a given position, the listening level at which the different loudspeakers were set could then significantly vary. This observation confirms the fact that the room behavior depends on both the position and the frequency response of the source, as was already observed by Bech [4].

The important result regarding this analysis is that the loudspeakers were globally not listened to at different levels, the preferred level being the same for all loudspeakers. The procedure Pr_3 enabled each listener to set the level but in average it did not introduce loudness differences between loudspeakers. Preferential differences compared to Pr_2 cannot be charged to non-matched listening levels which have already proven to affect the subjective assessments [16]. The observation of the level setting during a trial also proved that the listeners followed the instructions and compared the two stimuli once the two levels were set. It might nevertheless be argued that this operation could give some hints about the loudspeaker behavior at various levels and finally influence the final judgment. On the other hand, a t-test showed that the average durations of procedures 2 and 3 (respectively 16'33" and 19'28") were not statistically different ($t(47) = -1.63;\ p = 0.11$). The slightly higher discrimination in Pr_3 cannot in any case be caused by the fact that the listeners spent more time on this procedure as the level setting process did not significantly increase the session duration. The listeners' settings proved to be

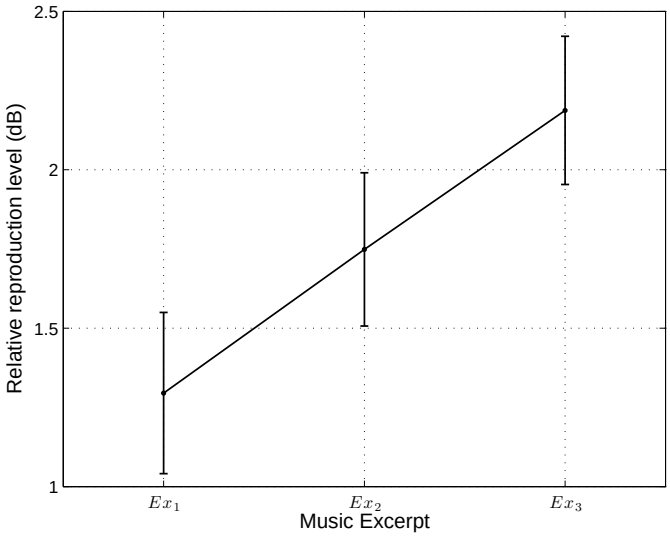


Figure 9: Mean level settings for the 3 excerpts, within their 95% confidence intervals.

significantly higher than the ones initially chosen by the expert listeners (see Fig. 9). The relative increase was quite moderate (2.2 dB at the most) and should not change the loudspeaker behavior in a highly significant way. This procedural difference caused small but significant differences in the sound quality ratings obtained in Pr_3 compared to Pr_2 (see Fig. 5). As a result, the setting process enabled the listeners to select a comfortable level that was significantly different from the preset one (i.e. moderately higher) and slightly enhanced the discrimination between loudspeakers.

5. Conclusion

This paper dealt with the comparison of three different procedures designed to assess loudspeaker sound quality. The listeners were proposed successive or alternate listening, for the latter case the reproduction level was matched beforehand or set by the subject himself. These assessment procedures were compared on the basis of the sound quality ratings obtained by four different loudspeakers in each session and the main result of this comparative study is that the subjective assessments were stable over the different procedures. For three different music excerpts, the three designed

procedures gave consistent results, although it can be argued that intrinsically different stimuli were used, namely short and long ones. As a rule, one of the loudspeakers obtained significantly lower quality ratings than the other ones. The three other loudspeakers obtained very similar quality ratings when considering all excerpts and assessment procedures.

Nevertheless, the listener's ability to separate between loudspeakers perceived as qualitatively very close proved to be dependent on the excerpt and procedure used to assess sound quality. The procedures under test exhibited thus differences in their discrimination power among the assessed loudspeakers using three different music excerpts. It appeared that the procedures 2 and 3, that were assumed to be more natural than a classical AB comparison of short excerpts, enabled to obtain significantly different ratings whereas statistically equivalent ones were obtained in procedure 1. Even though the comparison task was supposed to be more reliable using short excerpts, it appeared that such stimuli did not enable to separate between loudspeakers having very similar sound qualities. Nevertheless, this procedure proved to have a significantly shorter duration. Under certain conditions, the longer excerpts enabled the listeners to give finer quality ratings, especially when they were allowed to set the reproduction level. The listeners' settings proved to be slightly but significantly higher than the listening levels that were initially chosen by the expert listeners.

Acknowledgements

The authors wish to thank the staff and students of the "Image & Son" department from the University of Brest for participating in this experiment.

The authors would also like to thank Pierre-Yves Diquelou from Cabasse Acoustic Center (Plouzané, France) for his help in selecting the excerpts and matching the systems.

References

[1] U. Jekosch, Basic concepts and terms of "quality", reconsidered in the context of product-sound quality, Acta Acust United Ac 90 (6) (2004) 999–1006.

[2] S. Olive, P. Schuck, M. Sally, S. Bonneville, The effects of loudspeaker placement on listener preference ratings, J Audio Eng Soc 42 (9) (1994) 651–669.

[3] A. Gabrielsson, H. Sjögren, Perceived sound quality of sound-reproducing systems, J Acoust Soc Am 65 (4) (1979) 1019–1033.

[4] S. Bech, Perception of timbre of reproduced sound in small rooms: Influence of room and loudspeaker position, J Audio Eng Soc 42 (12) (1994) 999–1007.

[5] F. Toole, Sound reproduction – The acoustics and pyschoacoustics of loudspeakers and rooms, Focal Press, Oxford, UK, 2008.

[6] AES20–1996 (1996, reaffirmed 2007), AES recommended practice for professional audio – Subjective evaluation of loudspeakers, Journal of the Audio Engineering Society **44**, 382–400, Audio Engineering Society, New York City, NY, USA.

[7] IEC 60268–13 (1998), Sound system equipment – Part 13: Listening tests on loudspeakers, International Electrotechnical Commission, Geneva, Switzerland.

[8] ITU–R BS.1284–1 (2003), General methods for the subjective assessment of sound quality, international Telecommunications Union, Geneva, Switzerland.

[9] F. Toole, Subjective measurements of loudspeaker sound quality and listener performance, J Audio Eng Soc 33 (1/2) (1985) 2–32.

[10] E. Zwicker, H. Fastl, Psychoacoustics – Facts and models, 2nd Edition, Springer, New York City, NY, USA, 1998.

[11] S. Olive, A multiple regression model for predicting loudspeaker preference using objective measurements: Part 1 – Listening test results, in: Proceedings of the AES 116$^{\text{th}}$ convention, 2004, paper no. 6113.

[12] ITU–R BS.1534–1 (2003), Method for the subjective assessment of intermediate quality levels of coding systems, international Telecommunications Union, Geneva, Switzerland.

[13] S. Bertet, J.Daniel, E. Parizet, O. Warusfel, Influence of microphone and loudspeaker setup on perceived Higher Order Ambisonics reproduced sound field, in: Proceedings of Ambisonics Symposium, 2009.

[14] J. Käsbach, S. Favrot, J. Buchholz, Evaluation of a mixed-order planar and periphonic Ambisonics playback implementation, in: Proceedings of Forum Acusticum, 2011.

[15] M. Lavandier, P. Herzog, S. Meunier, Comparative measurements of loudspeakers in a listening situation, J Acoust Soc Am 123 (1) (2008) 77–87.

[16] A. Gabrielsson, U. Rosenberg, H. Sjögren, Judgments and dimension analyses of perceived sound quality of sound-reproducing systems, J Acoust Soc Am 55 (4) (1974) 854–861.

[17] B. Moore, B. Glasberg, T. Baer, A model for the prediction of thresholds, loudness, and partial loudness, J Audio Eng Soc 45 (4) (1997) 224–240.

[18] B. Glasberg, B. Moore, A model of loudness applicable to time-varying sounds, J Audio Eng Soc 50 (5) (2002) 331–342.

[19] J. Chalupper, H. Fastl, Dynamic loudness model (dlm) for normal and hearing-impaired listeners., Acta Acust United Ac 88 (3) (2002) 378–386.

[20] J. Rennies, J. Verhey, H. Fastl, Comparison of loudness models for time-varying sounds, Acta Acust United Ac 96 (2) (2010) 383–396.

[21] H. Møller, Fundamentals of binaural technology, Appl Acoust 36 (3–4) (1992) 171–218.

[22] S. Olive, T. Welti, W. Martens, Listener loudspeaker preference ratings obtained in situ match those obtained via a binaural room scanning measurement and playback system, in: Proceedings of the AES 122$^{\text{nd}}$ convention, 2007, paper no. 7034.

[23] T. Hiekkanen, A. Mäkivirtä, M. Karjalainen, Virtualized listening tests for loudspeakers, J Audio Eng Soc 57 (4) (2009) 237–251.

[24] S. Olive, B. Castro, F. Toole, A new laboratory for evaluating multichannel audio components and systems, in: Proceedings of the AES 105th convention, 1998, paper no. 4842.

[25] A. Watson, K. Attenborough, N. Heap, Assessing the subjective impact of loudspeaker response errors, Appl Acoust 41 (2) (1994) 157–168.

[26] A. Gabrielsson, B. Schenkman, B. Hagerman, The effects of different frequency responses on sound quality judgments and speech intelligibility, Hear Res 31 (1988) 166–177.

[27] F. Toole, S. Olive, The modification of timbre by resonances: Perception and measurement, J Audio Eng Soc 36 (3) (1988) 122–142.

[28] A. Gabrielson, B. Lindström, O. Till, Loudspeaker frequency response and perceived sound quality, J Acoust Soc Am 90 (2) (1991) 707–719.

[29] S. Zielinski, F. Rumsey, S. Bech, On some biases encountered in modern audio quality listening tests – A review, J Audio Eng Soc 56 (6) (2008) 427–451.

[30] M. Jason, Design considerations for loudspeaker preference experiments, J Audio Eng Soc 40 (12) (1992) 979–996.

[31] F. Toole, Subjective measurements of loudspeakers: A comparison of stereo and mono listening, in: Proceedings of the AES 74th convention, 1983, paper no. 2023.

[32] A. Devantier, S. Hess, S. Olive, Comparison of loudspeaker-room equalization preferences for multichannel, stereo, and mono reproductions: Are listeners more discriminating in mono?, in: Proceedings of the AES 124th convention, 2008, paper no. 7492.

[33] S. Choisel, F. Wickelmaier, Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference, J Acoust Soc Am 121 (1) (2007) 388–400.

[34] ITU–R BS.775–2 (2006), Multichannel stereophonic sound system with and without accompanying picture, international Telecommunications Union, Geneva, Switzerland.

[35] H. A. David, The method of paired comparisons, 2nd Edition, Griffin, London, UK, 1998.

[36] ISO 8586–2 (2008), Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 2: Experts, International Organization for Standardization, Geneva, Switzerland.

[37] E. Parizet, N. Hamzaoui, G. Sabatié, Comparison of some listening test methods: A case study, Acta Acust United Ac 91 (2) (2005) 356–364.