# Evaluation of Sound Quality, Boominess, and Boxiness in Small Rooms*

**ADAM WEISSER** AND JENS HOLGER RINDEL,** *AES Member*
(adam_weisser@fastmail.fm) (jhr@oersted.dtu.dk)

*Acoustic Technology, Technical University of Denmark, DK-2800 Lyngby, Denmark*

The acoustics of small rooms have been studied with emphasis on sound quality, boominess, and boxiness when the rooms are used for speech or music. Seven rooms with very different characteristics were used for the study. Subjective listening tests were made using binaural recordings of studio-produced music and anechoic speech. The test results were compared with a large number of objective acoustic parameters based on the frequency-dependent reverberation times (RT) measured in the rooms. This has led to the proposal of three new acoustic parameters, which have shown high correlation with the subjective ratings. The classical bass ratio definitions showed poor correlation with all subjective ratings. The overall sound quality ratings gave different results for speech and music. For speech the preferred mean RT should be as low as possible, whereas for music a preferred range of between 0.3 and 0.5 s was found.

## 0 INTRODUCTION

This experiment was set up in an attempt to investigate the influence of specific small-room acoustics on the listening experience in the room or, more precisely, on the perceived sound quality of the room. It emulates monophonic recorded music listening and quasi-live talker listening, compromising on a wrong directionality pattern of the latter.

The acoustics of small rooms are characterized by complicated modal acoustics, more dominant at higher frequencies than in large rooms, for which statistical acoustics approximations cannot be used. Wave phenomena such as diffraction cannot be neglected, and hence geometrical room acoustics are not good either. In addition, early reflections take place much sooner after the direct sound, and the whole reflection regime is compressed in time compared to large rooms. All these factors combined entail that many of the objective room parameters commonly used for large-room acoustical design are useless when dealing with small rooms (see, for instance, Kuttruff [1] and Vorländer [2]).

At very low frequency bands the reverberation time may vary significantly within the room volume, as its measurements occasionally represent some dominant modes, rather than a large ensemble average. Therefore few modal decay times are prominent in the mean reverberation time (RT) of low bands. However, we believe that whether the RT figure represents an accurate statistical average of a diffused field or a coarser average of more local modal decay times, it is still a valid measure, which can relate closely to various listening perceptions in the room.

A few more precise points of interest were defined at the onset of this project:

1) How well do the existing acoustical objectives describe, qualify, and quantify the sound quality in small rooms? Notably the reverberation time (RT), early decay time (EDT), and bass ratio (BR) were examined.

2) Is a flat curve of frequency-dependent RT preferable to a nonflat curve with a longer RT at low frequencies?

In order to test the preceding, it was decided to survey a variety of small rooms having different characteristics, which will give a spread of RT curves. A previously prepared program would be played monophonically via a high-fidelity loudspeaker and recorded binaurally with a dummy head in a few source–receiver positions within each room. The resulting library of dummy-head recordings would be the raw material for subjective listening tests, in the form of a sound-quality (SQ) rating scale test with the recorded material, using test subjects of various backgrounds. In addition, the test subjects would be asked to rate the recording “boominess” and “boxiness” in an attempt to correlate them later with longer RT values at low frequencies and perhaps with perceived coloration, although the latter was never quantified. An objective

*Manuscript received 2005 April 25; revised 2005 November 15 and 2006 March 20.

**Now with Oticon A/S, Smørum, Denmark.

room parameter that possibly correlates with boominess is the room bass ratio.

Since this is essentially a preference test, there is a risk of a large spread of results, which eventually will not be very telling. Obtaining consistent results from untrained listeners is by no means guaranteed, because of the material quantity, the definitions used for the rating tasks, and the subtleties involved, with which they will have to deal.

The source material was kept monophonic in order to remove stereo-imaging effects of the reproduced sounds, such as a distorting image shift. This controlled program material is not the usual listening material at home for most listeners. However, it accentuates some of the room acoustical features, which are likely to be dominant in stereophonic reproduction as well, namely, mean RT and room modes. The type of samples used and the loudspeaker directionality excluded any option for simulating the playing of live instruments.

## 1 SELECTED ROOMS

Seven small rooms were used to construct a recorded-sound library. The rooms are located in three different buildings with somewhat different construction standards. Two rooms in the Danish Radio (Talk Studio 8 and control room for Studio 3) are acoustically designed for professional use and fall into the category of acoustically "good" rooms, with relatively flat RT curves, irregular construction shapes to avoid strong standing waves, and so on. Besides these rooms, the only other designed room is an IEC standardized listening room whose use here coincides with its usual application.

The other four rooms were a meeting room, a classroom, a library, and a hearing-protector testing room. The latter was designed to have maximum RT, so it is considered an acoustically "bad" room for listening and was chosen as a reference room, representing a margin. All are box-shaped, but furnished and treated differently. These rooms were picked with the purpose of representing various typical environments.

The same setup was used for all rooms. Six measurement points (with three independent receiver positions for two source positions) were averaged for the final estimates of the room RT. Only in the library were eight positions measured. Software capable of maximum-length sequence (MLS) impulse measurements was employed, from which the RT and EDT were later extracted. In each measurement a 5.46-s MLS cycle was set to run ten times. The measurement procedure follows the ISO standard described in [3].

The background noise level in the rooms was also measured and computed according to both NR and RC noise ratings ([4], [5]). The level of the programs played had to be rather high to obtain a reasonable signal-to-noise ratio (SNR). It was around 80 dB A-weighted SPL.

All RT and EDT curves of the rooms are summarized in Figs. 1 and 2 along with the system minimum RT threshold, as was measured in an anechoic chamber. The standard [3] requires a minimum of 45-dB SNR for $T_{30}$ measurements and 35 dB for $T_{20}$. The classroom is the only room where it is lower than 40 dB below 63 Hz. Otherwise all rooms showed high SNRs above 50 Hz, which is the lowest frequency band used for the subsequent analysis, and the measurements comply with the standard [3]. The

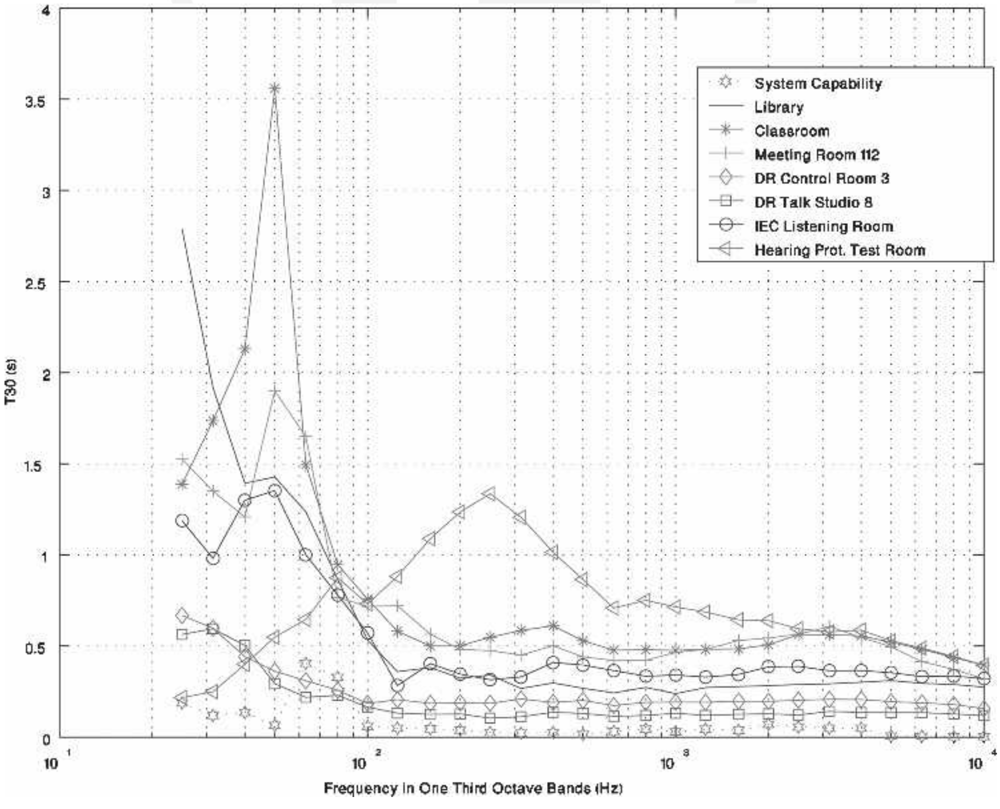

Fig. 1. Measured average $T_{30}$ of all rooms in experiment.

room dimensions, background noise, and materials are summarized in Tables 1, 2, and 3.

## 2 RECORDINGS

The recording chain used is illustrated in the block diagram in Fig. 3. It is described in detail in the following.

### 2.1 Recording Chain

All recordings were binaural. The source was monophonic and the recordings were two channel. A KEF reference 105.2 loudspeaker was used as the source. Occasionally it was placed on a loudspeaker stand, about 1 m in height. The loudspeaker has one separate enclosure for the woofer and another one for the midrange and the tweeter. The loudspeaker frequency response was measured in an anechoic chamber, and the relative narrow-band response was recorded using an FFT analyzer. The transfer function of the loudspeaker can hardly be considered flat, but it shows a ±3-dB deviation from flat response between 50 and 3000 Hz. Nonetheless its fidelity sounds satisfactory. Off axis the high frequencies above 2000 Hz drop slowly relative to the on-axis response, but the function's general shape is maintained. Since the low-frequency response is instrumental for the conclusiveness of the experiment with regard to low frequencies, this model was favored over some other smaller yet good loudspeakers whose bass responses are not low enough.

The loudspeaker was driven by a Labgruppen 300 amplifier, which was also used for the room acoustics measurements. An NAD CD player played the source material from a compilation CD.

### 2.2 Source Programs

Two programs were chosen for the rating experiment and a third one was used for the familiarization/training

Table 1. Principal room dimensions.

| Room | Length (m) | Width (m) | Height (m) | Volume ($m^3$) |
|---|---|---|---|---|
| Meeting | 6.23 | 4.18 + 0.40<br>(window niche) | 3.04 | 84.5 |
| Lecture | 9.45 | 6.27 + 0.40<br>(window niche) | 3.01 | 186.3 |
| Library | 9.45 | 6.27 + 0.40<br>(window niche) | 3.01 | 186.3 |
| Control | NA | N/A | NA | 100–110* |
| Talk studio | NA | N/A | NA | 80–85* |
| Hear. pro. test | 4.13 | 2.75 | 2.35 | 26.6 |
| IEC listening | 7.50 | 4.72 | 2.75 | 97.3 |

* Estimated.

Table 2. Room background noise ratings.*

| Room | NR | RC* |
|---|---|---|
| Meeting | 20 | 19 (N) |
| Lecture | 25 | 20.3 (N) |
| Library | 25 | 20.3 (N) |
| Control | 25 | 18 (R) |
| Talk studio | 25 | 18.3 (R) |
| Hear. pro. test | 15 | 10.3 (N) |
| IEC listening | 15 | 9.3 (N) |

* (N)—neutral noise spectrum; (R)—Rumbly spectrum.

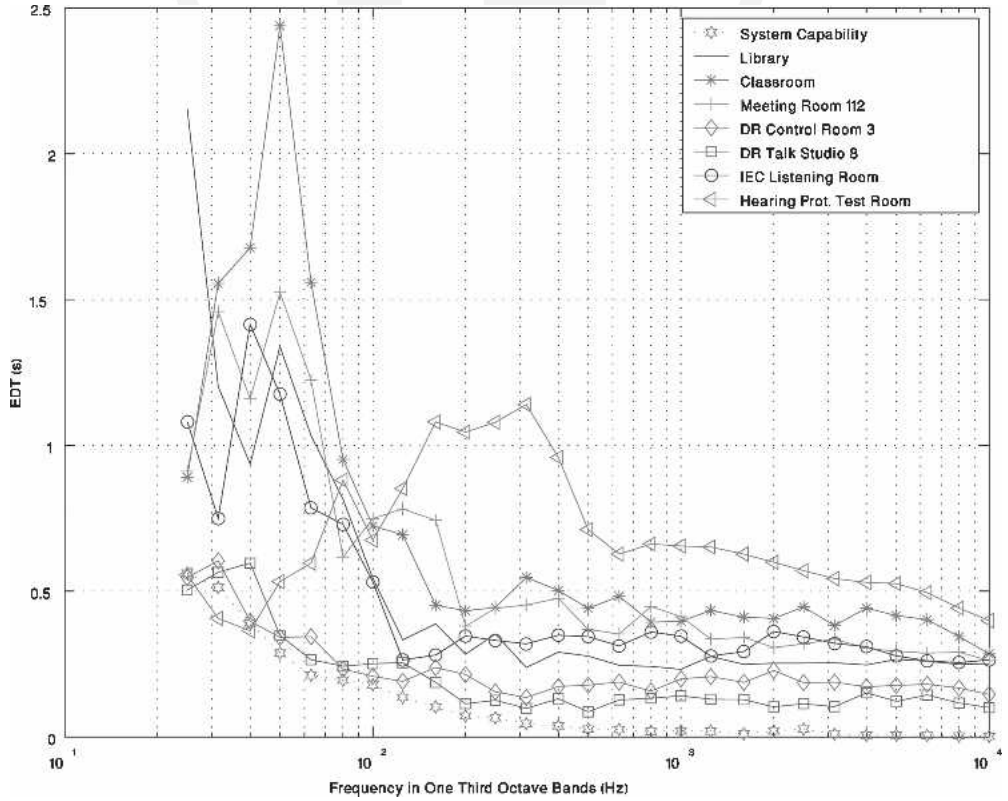

Fig. 2. Measured average early decay time (EDT) of all rooms in experiment.

round of the rating task. All excerpts were 20 s long. The three programs were as follows:

1) *Male anechoic speech* (B&O CD, Archimedes project 1992).

2) "*The Audience,*" Matthew Herbert, taken from "Bodily Functions" (2001), !K7. Electronic music with human noise samples and a female (soprano) singer. It also contains synthesized ambient noise.

3) "*Spell,*" Jimi Tenor, taken from "Out of Nowhere," Warp, 2000. Brass, wind, and string sections, drums, bass, and electric guitar.

Note that the subjects performed a different experiment earlier on (not mentioned in the text), which used similar or identical samples and can be viewed as another form of familiarization round.

The music samples were chosen first according to their original recording fidelity, which should be the highest available. Then the excerpts should still exhibit good quality in any one of the monophonic reproduction possibilities: right, left, or mono. The latter was never used. The relative power spectrum of the samples is shown in Fig. 4. The most important question was how well a certain sample can be used in order to pinpoint the various room acoustics features.

Two more aspects were considered before and after the tracks were chosen. First, the music should not be too

Table 3. Room materials.

| Room | Floor | Walls | Ceiling |
|---|---|---|---|
| Meeting room | 22-mm parquet on joists with a cavity of around 10-mm air. | Brick walls with one gypsum wall. Windows form another wall starting from about 0.90 m. | Suspended ceiling with mineral wool layer and air cavity of about 0.40 m below concrete roof. |
| Lecture room | Same as meeting room. | Same as meeting room, but without gypsum wall. | Same as meeting room. |
| Library | Same as meeting room. | Same as meeting room, but without gypsum wall. | Same as meeting room. |
| Control room | Uncertain materials and dimensions. | Uncertain materials and dimensions. | Uncertain materials and dimensions. |
| Talk studio | Carpet on concrete floor. | 40-mm mineral wool in steel frames covered with fabric with a 13-mm gypsum board. | Perforated standard metal cassette ceiling. |
| Hearing protector test room | Floating floor on mineral wool springs, no joists. | Three gypsum board layers over mineral wool and air cavity. | Three gypsum board layers over mineral wool and air cavity. |
| IEC listening room | Same as meeting room, carpeted. | Uncertain. | Uncertain. |

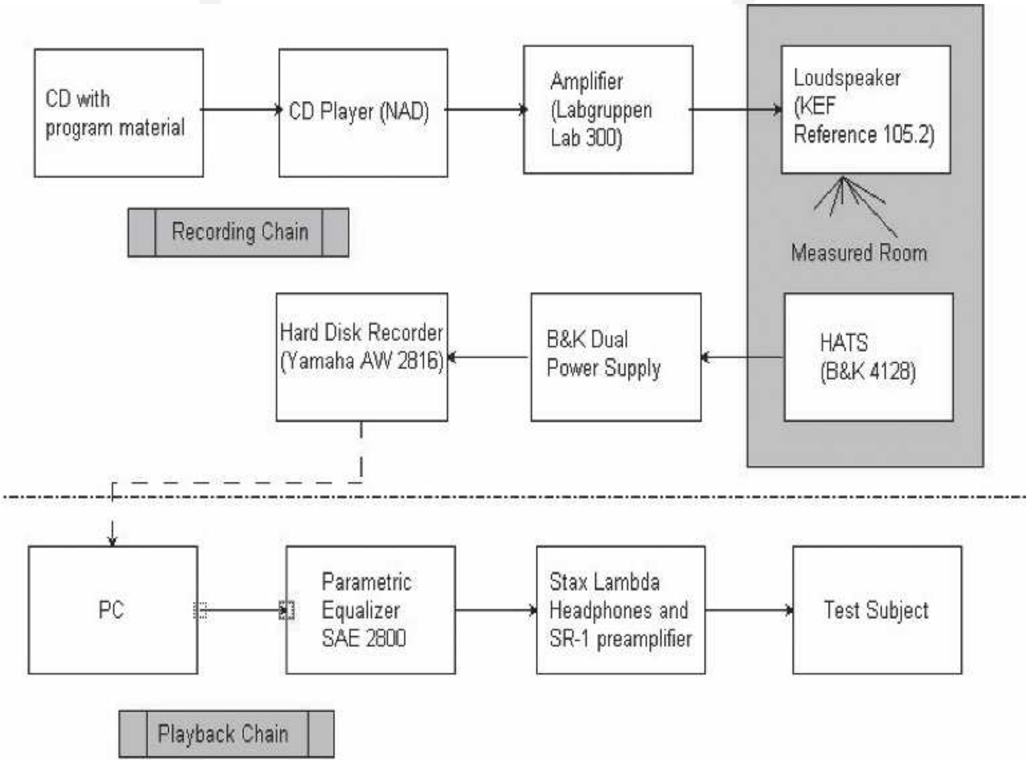
CD with program material
CD Player (NAD)
Amplifier (Labgruppen Lab 300)
Loudspeaker (KEF Reference 105.2)
Recording Chain
Measured Room
Hard Disk Recorder (Yamaha AW 2816)
B&K Dual Power Supply
HATS (B&K 4128)
PC
Parametric Equalizer SAE 2800
Stax Lambda Headphones and SR-1 preamplifier
Test Subject
Playback Chain

Fig. 3. Block diagram of recording and playback chains.

appealing subjectively and should not arouse or annoy. Second, the tracks should not be familiar to the subjects. Both considerations are necessary to reduce unwanted bias.

The male speech recordings were chosen for speech evaluation, for their wider spectral content compared to female speech.

The Herbert track was picked for the music signal evaluations. The reason was that its percussive nature and its many details facilitated the impression of the room acoustics with regard to the detail resolution, that is, the effects of RT on single events and on continuous signals at the same time. All in all, its instrumentation is simple enough for a listener not to lose track in a sea of sound. The utilization of electronic music for the test was, in a way, a shot in the dark. It contains no familiar instruments (apart from a female vocalist), but only sampled or synthesized sounds, which can be referenced only to themselves and not to any idea that may preexist in the listener's mind. The original recording is a more popular electronic music production, with overall low reverberation and rather percussive in nature.

### 2.3 Recording Microphone

In order to obtain more realistic simulations of the small-room acoustics, the recordings were binaural, using a dummy head. The binaural technique captures the entire auditory signal arriving at each ear drum if a listener were situated in place of the dummy. A successful binaural playback would emulate the original sonic environment, complete with spatial information.

A B&K head and torso simulator (HATS) type 4128 was used as dummy head for all binaural recordings. There are no ear canals in the ear molds, so no correction needs to be made subsequently for their effect. The HATS is built as a human body half with a torso, wearing a vest.

Initial tests in the anechoic chamber showed very good sound fidelity with mixed localization performance, depending on the direction of incidence. It compared with a similar performance figure, reported in Møller et al. [6] specifically for the 4128 HATS. To overcome the problem of increased mislocalization on the median plane (front–above–back) of the binaural replication most of the recordings in the sound library were not recorded on axis, but off axis to some degree. These areas are easier to localize in the playback. The price for this feature is that the recordings have varying left–right channel "balance," which the subjects must disregard.

### 2.4 Recorder

A hard disk recorder (HDR) by Yamaha, type AW 2816, was used because of the convenience of computerized transfer of recorded samples to WAV format or directly onto an audio CD. It includes two analog microphone inputs, enough for the HATS, which are then converted to digital inside the machine.

### 2.5 Recording Positions

In order to represent the rooms better in the recorded library, each recording was made with three or four different positions of source and receiver. Each position had different frequency response characteristics, so that different source–receiver locations do not sound identical. Only one position per room was chosen to represent a room in the rating test, thus eliminating one factor from the analysis.

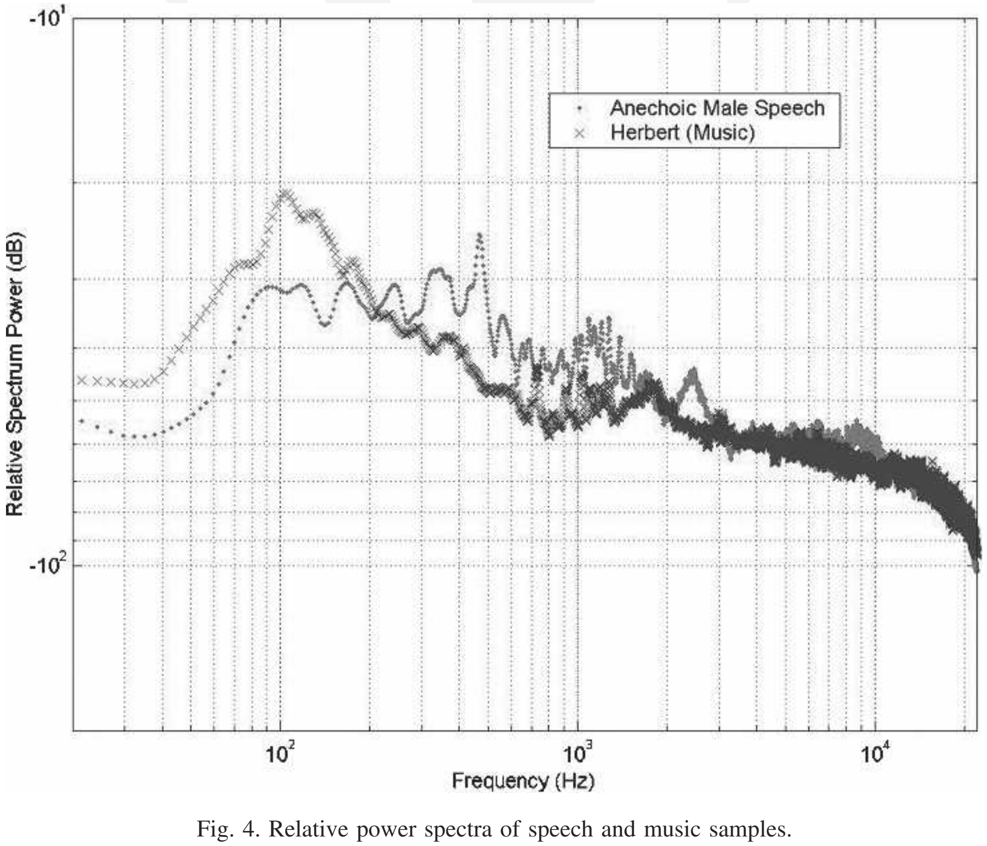

Fig. 4. Relative power spectra of speech and music samples.

## 3 PLAYBACK CHAIN

The playback chain is shown in the bottom part of Fig. 3. The fidelity of the recording equipment is also relevant for the playback chain. The most complicated part is the binaural reproduction through headphones if it is to give an impression of a true virtual room.

### 3.1 Recorded Media, Playback Machine, and Amplifier

The samples were recorded digitally and thus do not degrade when manipulated and played back repeatedly. The playback quality should match the recorder playback capability, using a decent digital-to-analog converter. Samples are played as WAV format files from a PC and fed from the PC's sound card to the headphone preamplifier, through a parametric equalizer. Avance 97 was the standard sound card used for playback.

### 3.2 Headphones

The Stax Lambda headphones used throughout the experiment are able to produce accurate sound, which is closer to the original source sound. These circumaural electrostatic headphones include a special class A electron tube preamplifier (Stax SR-1), which also provides the polarization voltage needed for the headphones to operate. Møller et al. [7] measured 14 types of headphones and compared their headphone transfer functions (PTFs), that is, the individual frequency response that arrives at the listener's eardrum. The Stax Lambda showed a relatively small spread and variation between test subjects. Even so, it is clear that the PTFs are by no means identical for all test subjects, and as the frequency increases so do the differences between subjects.

There were two immediate problems detected in the pilot test recordings. The first was overemphasis of the small-room acoustics. Namely, a stronger impression of the room acoustics was conveyed through the recordings than through listening to an identical live presentation, as if the room acoustics were more dominant than they really are. The second was an in-head sense of a median plane (front–back) for the recording in the anechoic chamber. The two problems are likely to stem from the same source, namely, the nonindividual head transfer function used with the HATS and the nonequalized headphone transfer function. Only the latter could be addressed with the available means.

### 3.3 Equalizer

Even with high-quality headphones there are still significant differences in the way different subjects perceive the binaural recordings spatially. It is generally established that equalization of the headphone-to-microphone transfer function to obtain a flat response is the minimum needed to obtain so-called nonindividualized reproduction. Individualized reproduction occurs if such equalization is done individually, using the subject's ears [8].

For the scope of this work it was decided instead to make use of existing analog filters and equalize mainly the low and midrange frequencies of interest. The headphones were equalized only up to a few kHz. Another reason was that correction for high frequencies using this method introduces substantial noise (hiss). It is unclear how important the change in the original phase in the recordings is, but it is assumed negligible, so that the two channels are not equalized exactly the same because of slight differences in the headphone–preamplifier channels. SAE 2800 parametric equalizer resonant filters were used to correct for the transfer functions, which was monitored in real time with a dual-channel signal analyzer.

## 4 EXPERIMENT

The entire test was automated and computerized using MATLAB codes. Headphone playback permitted having the test in a normal (nonlistening) quiet room, where the subject is left with the system. Despite haphazard noise events, the listeners were able to concentrate well and did not report any noise-related problems in doing the tests. The subjects were given no time constraints. Usual running times varied from an hour and a half to two and a half hours (including tests not described in this paper). Subjects were instructed to take breaks between the two parts of each experiment and between experiments in order to avoid fatigue.

Due to the overall multitude of samples used in this experiment and two others that were used during this listening test, the samples did not have exactly the same perceived loudness. Sometimes there was a minute difference between samples, and subjects were instructed to adjust the volume slightly when necessary, using the Stax preamplifier's volume knob. Nevertheless, in most cases the listeners did not change the presentation level but kept it constant at a comfortable setting throughout the rating test.

The binaural aspect of the recordings was mentioned as well. Subjects were told to notice that it may not always work (in-head virtual sound source) and that experiencing the spatial effects is not essential for accomplishment of the task.

Finally the subjects were guided to try and ignore the differences between sample volume and channel balance—an artifact of the particular binaural recording setting used (see Section 2.3)—when making decisions.

### 4.1 Sound Quality, Boominess, and Boxiness Rating Test

It was impossible to cover all positions within a room in all programs. Therefore two representative tracks were chosen for the rating test. One recording position was chosen for each room, giving a total of 14 samples that the subjects had to evaluate. The subjects were requested to note their impressions of the overall sound quality, boominess, and boxiness of the samples according to the following written guidelines:

- "Overall Sound Quality"—A general measure of your sensation and satisfaction of the recorded sample quality, which might be combined from several things such as: detail resolution, annoyance / pleasure of the sound (not the content though!), natural / artificial sound, and anything else you may see relevant.

- "Boominess"—A boomy recording can be defined as having an excessive bass and/or a bass lacking definition, which is "smeared" over time (as opposed to being "punchy"). The general feeling is imitative of a boom.
- "Boxiness"—A boxy sound can be taken quite literally as sounding in a well-defined box. Think of sound heard in a typical bathroom or at times when you talk inside a closet (if you hid there as a child), as typical illustrations.

The instructions were repeated verbally and discussed in order to make sure that they were understood clearly. Subjects were encouraged to stick their heads into a wooden closet available in the room and say something to get an illustration of a severe boxy sound.

For familiarization with the rating scales used and the degree of variation between samples, three introductory examples were given for rating of another recording in three very different rooms: a talk studio, a lecture room, and a hearing-protector testing room. Samples could be played back as many times as the subjects wished before the three ratings were given successively.

The order of presentation was randomized using a Latin square scheme, and after a short break the 14 samples were presented again in a different order. The following rating scale was used for overall sound quality evaluations:

1. Intolerable
2. Very annoying
3. Unpleasant
4. Not so good
5. Acceptable
6. Decent
7. Good
8. Very good
9. Excellent

The scale for boominess ratings was:

1. Very thin, very hollow
2. Thin, hollow
3. Slightly thin, hollow
4. Balanced
5. Slightly boomy
6. Boomy
7. Very emphasized

Both scales are bipolar and could have been represented by a negative–positive scale. It was chosen not to do so to avoid unnecessary confusion.

The boxiness scale is unipolar:

1. Unnoticeable
2. Barely audible
3. Distinct yet not dominant
4. Dominant
5. Very dominant

## 4.2 General Presentation Details

### *4.2.1 Test Subjects*

A total of 18 subjects were tested on the three experiments, including one pilot test subject. After the initial pilot tests only minute modifications were made, and so the pilot subject's data were also used in the final analysis. Test subjects were between 20 to 40 years old and all were tested for normal hearing either recently or prior to the tests, according to ISO 389 for hearing threshold measurements. The test subjects included 14 males and 4 females, of whom 13 were students or staff at the acoustics department and 5 were completely untrained subjects (as far as room acoustics are concerned; three had some music background).

### *4.2.2 Presentation Order*

Three listening tests were performed by each subject. The middle one was the rating test in question. The other two are not described in this paper, but they followed identical general procedures and used the same sample library. It was hoped that the first experiment would help the subjects to become further familiarized with the samples and the range of room acoustics involved.

### *4.2.3 Presentation Level*

Subjects were instructed to listen to the samples at a comfortable level and thus, if samples were somewhat different in level, to compensate for that by using the volume knob of the Stax preamplifier. Still, most subjects tended to keep a more or less fixed level. The average presentation level was measured using a B&K artificial ear. The ear was connected to a calibrated measuring amplifier (B&K 2607, calibrated at 1000 Hz using a sound level calibrator B&K type 4230). The A-weighted level of the samples varied between 70 and 80 dB, depending on the program material, the particular recording, and the individual volume setting.

There was a concern that variable presentation levels will unwantedly influence the bass perception of a subject because of the compression of bass dynamics in loudness perception. Equal loudness contour levels are strongly dependent on the sound pressure level. At lower levels the bass is perceived to be relatively weak, compared to higher frequencies. The higher the sound pressure level goes, the more the bass response evens out with regard to higher frequencies. At levels corresponding to our presentation levels (around 50–60 dB SPL at 1 kHz), the difference between the curves is not as pronounced as for lower sound pressure levels. However, the curves were measured for pure tones, and the application to complex signals is not straightforward.

Two reasons were given for not fixing the presentation levels. First it was acknowledged that the sample loudness levels were not equal. Time constraints and the amount of samples did not allow for a comprehensive normalization. Second, the ability to set the volume is optional in domestic situations, and so a subject who is to give a critical opinion should do so under his or her preferred conditions.

There is a justified concern that the experiment is not well-controlled if an active modification of the output level is allowed. However, the presentation level was relatively well maintained, at around 75 dB A-weighted, where from a single loudness contour perception significant variations would not be expected. In case that modifications of the presentation level would affect the ratings after all through different bass perception sensitivity, they would enter the total random error of the test, so that the overall results and their significance will be decreased. The main concern in the variable presentation level was the boominess ratings. That question will be examined later in the analysis.

### 4.2.4 Binaural Reproduction

As was discussed before, much effort was put into binaural reproduction, which is to be as authentic as possible in relation to the original or, alternately, should convey a realistic sensation to the listener. From a haphazard sampling of the listeners it seems that all subjects experienced at times a very realistic spatial image, whereas in other cases they perceived an in-head source location. No specific samples were noted for either, as no comprehensive and systematic test was done in that regard. It seems, however, reasonable to assume that where the original source was located off axis, the replication was more realistic. That effect of unbalanced recordings was at times confusing for subjects, despite the previous training.

## 5 RESULTS AND ANALYSIS

### 5.1 Task Accomplishment

Despite the general abstract definitions that had to be rated, subjects generally fared well in understanding their meaning and subsequently in rating them. It was noticed that acousticians tended to overanalyze the terms and situations used much more than untrained subjects. In a couple of instances acousticians misconceived a term and used a preconceived notion for it. Thus two subjects' boxiness ratings were not used.

In addition, one of these subjects' dislike for the musical piece to be rated ("Herbert") was so severe that his overall sound quality rating was useless, as he looked for the rooms where the most annoying features of the music, in his opinion, were totally eliminated by masking by a long reverberant sound. Another subject admitted that he systematically rated the music lower than the speech, because of his dislike for it. His remark had later eased the verification of the halo error (stimulus effect) correction, which was performed on the data (see Section 5.2.2).

At least two subjects reported that their ratings of the overall sound quality of the speech samples were directly related to their impression from the speech intelligibility, which was always rather high. Their ratings were used anyhow since their interpretation is seen as a legitimate element of sound quality.

Some subjects complained about an initial difficulty with rate boominess, while others had difficulties with boxiness. A few subjects reported a mistyping error, which yielded an unintentional rating. Unfortunately these random errors could not be corrected, and they contribute to the total random error in the measurements.

### 5.2 Preparatory Analysis

The rating data from the listening test are composed of four factors: rater, room, program, and parameter. The analysis began by analyzing each parameter separately. A further factor was the double rating of each room by each rater due to the use of a test and retest structure, referred to in the text as part 1 and part 2 of the results, respectively. The parts were examined individually, but were also combined to increase the significance of some of the final results.

The statistical analytical procedures are similar for all three rating scales. The analysis includes the following stages: a comprehensive four-way analysis of variance (ANOVA) for each parameter and statistical model establishment; halo effect (stimulus) error correction; individual linear transformation of ratings to conform for a regular scale with invariable dispersion between raters, reestablishment of the statistical models, and derivation of all means per room and program; correlations and connections to room acoustic parameters.

Data obtained from the individual tests contained a file for each of the three ratings. The data were later manipulated separately per program. Means were further obtained for each program and were averaged over parts, and for total rating they were also averaged over programs.

### 5.2.1 Analysis of Variance

An overview of all the sound quality (SQ) rating data collected is best obtained using a four-way analysis of variance, which takes into account all factors. The model used also computes the second-order interactions between factors. Not all interactions are of interest, as they do not always make sense, and yet they are all presented in Table 4, the ANOVA[1] summary table.

The only insignificant factor (where Prob > 0.05) in the experiment is the test part, including second-order interaction terms. It means that the average ratings are stable between the two parts, as can be double-checked easily by calculating the reliability of the test. The correlation between the means of the rooms in each test part gives the reliability of the test, when it is structured in the test–retest

Table 4. Four-way ANOVA summary table of overall sound quality ratings.

| Source | Sum Sq. | df | Mean Sq. | *F* | Prob>*F* |
|---|---|---|---|---|---|
| Rater | 278.29 | 17 | 16.3697 | 12.5 | 0 |
| Room | 335.14 | 6 | 55.8565 | 42.64 | 0 |
| Program | 38.89 | 1 | 38.8889 | 29.69 | 0 |
| Part No. | 0.07 | 1 | 0.0714 | 0.05 | 0.8155 |
| Rater*Room | 286.58 | 102 | 2.8096 | 2.14 | 0 |
| Rater*Program | 98.04 | 17 | 5.767 | 4.4 | 0 |
| Rater*Part No. | 15 | 17 | 0.8824 | 0.67 | 0.8288 |
| Room*Program | 125.47 | 6 | 20.912 | 15.96 | 0 |
| Room*Part No. | 5.57 | 6 | 0.9279 | 0.71 | 0.6431 |
| Program*Part No. | 0.51 | 1 | 0.50079 | 0.39 | 0.5339 |
| Error | 430.95 | 329 | 1.3099 | | |
| Total | 1614.5 | 503 | | | |

[1] The condition for ANOVA states that the data be normally distributed and homogeneous (in respect to their variance). Only the four-way ANOVAs tested positive for normality. The smaller tables contained too few samples with too small a variance. They are used here, though, by restricting their generalization.

method, or more accurately it gives the coefficient of stability. The total reliability was $r = 0.9664$ for the overall sound quality, $r = 0.9222$ for the boxiness rating, and a significantly lower figure of $r = 0.6121$ for boominess.

All other factors are significant and thus the analysis would be three-way only. Of all significant interaction terms, only the *Rater*Program* term is not obvious and is also undesired in the test. It points to a halo effect (stimulus) error, since subjects showed a biased response or a preference of a certain program over the other. Fortunately this particular error can be corrected for, as shown in Section 5.2.2. First we note the initial general statistical rating model,

$$\mathrm{SQ} = \mu_{\mathrm{SQ}} + X_{\mathrm{Rater}} + X_{\mathrm{Program}} + X_{\mathrm{Room}} + X_{\mathrm{Room*Rater}} + X_{\mathrm{Room*Program}} + X'_{\mathrm{Rater*Program}} + \varepsilon \quad (1)$$

where the total measured rating is composed of the mean ratings $\mu$ and the *X,* which are the effects of individual raters, programs, and rooms. The two room interaction terms were added to the table from observation. They account for a large fraction of the total variance. They mean that subjects rate different rooms differently and that the room rating is not invariable for the program played. $X'_{\mathrm{Rater*Program}}$ is the interaction term between raters and programs, which is the halo effect error. The rest of the error in the rating is encompassed by $\varepsilon$. The boxiness and boominess models (Tables 5 and 6) include one less term and are therefore simpler,

$$\mathrm{Boxiness} = \mu_{\mathrm{Box}} + X_{\mathrm{Rater}} + X_{\mathrm{Room}} + X_{\mathrm{Room*Rater}} + X_{\mathrm{Room*Program}} + X'_{\mathrm{Rater*Program}} + \varepsilon \quad (2)$$

$$\mathrm{Boominess} = \mu_{\mathrm{Boom}} + X_{\mathrm{Rater}} + X_{\mathrm{Room}} + X_{\mathrm{Room*Rater}} + X_{\mathrm{Room*Program}} + X'_{\mathrm{Rater*Program}} + \varepsilon \quad (3)$$

where the effect of the program alone does not play a role in either rating.

### 5.2.2 Correction of the Halo Effect Error

A short glimpse at the mean ratings for the program shows that most subjects favored the speech recordings in both parts, although only two raters admitted so. Few subjects showed remarkable unbiasedness between the programs. The rather cumbersome correction procedure follows closely Guilford [9] and is not accounted for here. Ratings for overall sound quality, boxiness, and boominess were assumed independent, and the procedure was repeated for each one, repeating for the additional test part factor. The correction was repeated, whenever the halo effect error showed greater than 5% significance: SQ (both parts), boominess (part 2), and boxiness (part 2).

### 5.2.3 Linear Transformation of the Ratings

Ratings by different individuals naturally have different means and dispersions. For instance, subject A may center his SQ rating around 5 and rate only between 4 to 6, whereas another centers hers around 6 and rates all between 4 to 8. As we are interested in the eventual relative rankings of the rooms, it is helpful to conform all ratings to a standard scale, with one mean and one dispersion. This ensures that the ratings of all subjects are equally weighted [9]. The ITU standard suggests a simple transformation to perform this normalization [10]:

$$Z_i = \frac{x_i - \bar{x}_i}{s_i} \cdot s_{\mathrm{t}} + \bar{x}_{\mathrm{t}} \quad (4)$$

where $Z_i$ is the normalized result, $x_i$ is the rating of subject *i,* $\bar{x}_i$ is the mean rating of subject *i,* $s_i$ is the subject's standard deviation, $s_{\mathrm{t}}$ is the total standard deviation for all subjects, and $\bar{x}_{\mathrm{t}}$ is the total mean.

### 5.2.4 ANOVA Revisited

After applying this normalization the four-way ANOVA is repeated. A few changes are apparent in the revised data. The interaction between programs and raters has completely disappeared due to the halo error correction, and its probability is artificially brought to 1—no interaction. The normalization of the ratings reduced the overall variance—first by diminishing the variance over raters and then by doing the same for the other two interaction terms with raters. Similar data were obtained for boxiness and boominess.

As a formality, the superfluous terms with little contribution to the variance or those with unwanted effects can be excluded from the final ANOVA models and their small contributions therefore transferred into $\varepsilon$. Table 7 is an example of a reduced ANOVA table for the sound

Table 5. Four-way ANOVA summary table of boxiness ratings.

| Source | Sum Sq. | df | Mean Sq. | *F* | Prob>*F* |
|---|---|---|---|---|---|
| Rater | 38.912 | 16 | 2.432 | 3.49 | 0 |
| Room | 137.513 | 6 | 22.9188 | 32.91 | 0 |
| Program | 2.288 | 1 | 2.2878 | 3.29 | 0.0709 |
| Part No. | 0.254 | 1 | 0.2542 | 0.37 | 0.5462 |
| Rater*Room | 210.059 | 96 | 2.1881 | 3.14 | 0 |
| Rater*Program | 37.248 | 16 | 2.328 | 3.34 | 0 |
| Rater*Part No. | 12.139 | 16 | 0.7587 | 1.09 | 0.364 |
| Room*Program | 12.992 | 6 | 2.1653 | 3.11 | 0.0057 |
| Room*Part No. | 1.966 | 6 | 0.3277 | 0.47 | 0.83 |
| Program*Part No. | 0.002 | 1 | 0.0021 | 0 | 0.9562 |
| Error | 215.861 | 310 | 0.6963 | | |
| Total | 669.233 | 475 | | | |

Table 6. Four-way ANOVA summary table of boominess ratings.

| Source | Sum Sq. | df | Mean Sq. | *F* | Prob>*F* |
|---|---|---|---|---|---|
| Rater | 61.161 | 17 | 3.5977 | 4.42 | 0 |
| Room | 61.548 | 6 | 10.2579 | 12.6 | 0 |
| Program | 1.05 | 1 | 1.0496 | 1.29 | 0.257 |
| Part No. | 0.018 | 1 | 0.0179 | 0.02 | 0.8824 |
| Rater*Room | 126.381 | 102 | 1.239 | 1.52 | 0.0031 |
| Rater*Program | 43.558 | 17 | 2.5622 | 3.15 | 0 |
| Rater*Part No. | 29.732 | 17 | 1.7489 | 2.15 | 0.0055 |
| Room*Program | 17.27 | 6 | 2.8783 | 3.54 | 0.0021 |
| Room*Part No. | 15.246 | 6 | 2.541 | 3.12 | 0.005 |
| Program*Part No. | 0.002 | 1 | 0.002 | 0 | 0.9607 |
| Error | 267.875 | 329 | 0.8142 | | |
| Total | 623.839 | 503 | | | |

quality. The reduced SQ, boxiness, and boominess statistical models can be rewritten, respectively, as

$$\mathrm{SQ} = \mu_{\mathrm{SQ}} + X_{\mathrm{Rater}} + X_{\mathrm{Program}} + X_{\mathrm{Room}} + X_{\mathrm{Room*Rater}} + X_{\mathrm{Room*Program}} + \varepsilon \tag{5}$$

$$\mathrm{Boxiness} = \mu_{\mathrm{Box}} + X_{\mathrm{Room}} + X_{\mathrm{Room*Program}} + \varepsilon \tag{6}$$

$$\mathrm{Boominess} = \mu_{\mathrm{Boom}} + X_{\mathrm{Room}} + X_{\mathrm{Room*Program}} + \varepsilon. \tag{7}$$

In the two latter models the rater effect was omitted, despite testing significant, as it displayed a very small contribution to the total variance. As for boominess, the interaction term between rooms and parts adds a large fraction to the total variance. It reconfirms the relatively low reliability in the test–retest of the boominess.

## 5.3 Ranking

As the corrections described in the preceding are linear and were performed with no other types of error correction, the resultant ranking is unaffected by them. However, since the variance of the means is smaller, the confidence intervals are decreased, and the significance of the means with regard to each other improves. The ranking is determined by the output plot of MATLAB's one-way ANOVA Multcompare function, which illustrates the 95% confidence interval for each room with respect to all other rooms. Naturally, the more ratings are joined together, the more the relative ranks will be significant.

## 5.4 Correlation to Room Acoustical Data and New Definitions

After having obtained all the means, a basic examination can reveal whether there are any significant intercorrelations between the three rated parameters and correlations between them and the room acoustical parameters. The significance of the correlation coefficient was determined by the one-tailed *t* test. The inspection highlights are summarized in Table 8. The basic quantities that were inspected are mean RT, mean EDT, bass ratio (BR), and room volume. However, the definition of the bass ratio was tweaked and optimized to have the highest correlation with the measurements, as the traditional (large-room) definition showed very poor correlation. The BR is defined as [11].

$$\mathrm{BR} = \frac{T_{60}(125\ \mathrm{Hz}) + T_{60}(250\ \mathrm{Hz})}{T_{60}(500\ \mathrm{Hz}) + T_{60}(1000\ \mathrm{Hz})} \tag{8}$$

where the octave-band $T_{60}$ in the numerator are sometimes replaced with lower bands of 63 and 125 Hz. Only the latter is shown in Table 9, as the higher band BR shows even poorer correlation. The highest correlation was achieved with the sound quality in the music ratings ($r = 0.7399$), and the correlation with boominess was slightly lower.

Table 7. Four-way ANOVA summary table of overall sound quality ratings after halo effect correction and rating normalization.

| Source | Sum Sq. | df | Mean Sq. | *F* | Prob>*F* |
|---|---|---|---|---|---|
| Rater | 29.135 | 17 | 1.7138 | 12.37 | 0 |
| Room | 335.139 | 6 | 55.8565 | 403.21 | 0 |
| Program | 39 | 1 | 38.8889 | 280.73 | 0 |
| Rater*Room | 28.059 | 102 | 0.2751 | 1.99 | 0 |
| Room*Program | 125.472 | 6 | 20.912 | 150.96 | 0 |
| Error | 51.394 | 371 | 0.1385 | | |
| Total | 608.088 | 503 | | | |

$T_{30}$ is the mean of the one-third-octave band data measured between 200 and 4000 Hz. The EDT is calculated from the same bands, taking into account only the initial 10-dB decay. Another quantity examined is the recommended mean RT, appearing as $T_{\mathrm{m}}$ in the ITU and EBU standards for listening rooms [10], [12]. Here the ratio between the actual RT and the recommendation is examined,

$$\frac{T_{30}}{T_{\mathrm{m}}} = \frac{T_{30}}{0.25(V/V_0)^{1/3}} \propto \frac{T_{30}}{V^{1/3}} \tag{9}$$

where $V_0$ is a reference volume of 100 m$^3$.

An improvement in the correlations is shown where using several new specially optimized quantities. The small-room bass ratio (SBR) was defined using the one-third-octave $T_{30}$ values,

$$\mathrm{SBR} = 10 \log\left[\frac{T_{30}(63\ \mathrm{Hz}) + T_{30}(80\ \mathrm{Hz})}{T_{30}(250\ \mathrm{Hz}) + T_{30}(315\ \mathrm{Hz})}\right]\ \mathrm{dB} \tag{10}$$

Table 8. Highest and most significant correlations between all rating data and acoustical parameters in experiments for music, speech, and both averaged.

| | Correlation *r* | One-Tailed<br>*t* test* | *P* |
|---|---|---|---|
| *Music* | | | |
| SQ–SBR† | 0.8359 | 3.40 | 0.01 |
| SQ–BR | 0.7399 | 2.46 | 0.03 |
| Boxiness–SBR | –0.8872 | –4.30 | <0.005 |
| Boxiness–SEBR | –0.8712 | –3.97 | <0.006 |
| Boxiness–T$_{30}$ | 0.8084 | 3.07 | 0.012 |
| Boxiness–EDT | 0.8484 | 3.58 | 0.009 |
| BR–LHR | 0.8262 | 3.28 | 0.014 |
| Boominess–LHR | 0.8797 | 4.14 | <0.005 |
| Boxiness–$T_{30}/T_{\mathrm{m}}$ | 0.8537 | 3.66 | 0.008 |
| *Speech* | | | |
| SQ–SEBR | 0.8996 | 4.61 | <0.005 |
| SQ–EDT | –0.8956 | –4.50 | <0.006 |
| SQ–$T_{30}$ | –0.9139 | –5.03 | <0.005 |
| Boxiness–$T_{30}$ | 0.8625 | 3.81 | 0.006 |
| Boxiness–EDT | 0.8594 | 3.76 | 0.006 |
| Boxiness–SEBR | –0.8916 | –4.40 | <0.005 |
| Boxiness–SQ | –0.9611 | –7.78 | <0.005 |
| Boominess–LHR | 0.6703 | 2.02 | 0.05 |
| SQ–$T_{30}/T_{\mathrm{m}}$ | –0.8511 | –3.75 | 0.007 |
| Boxiness-$T_{30}/T_{\mathrm{m}}$ | 0.8146 | 3.14 | 0.013 |
| *Total* | | | |
| SQ–boxiness | –0.9188 | –5.20 | <0.005 |
| SQ–SEBR | 0.9009 | 4.64 | <0.005 |
| SQ–SBR | 0.8923 | 4.40 | <0.006 |
| SQ–$T_{30}$ | –0.7882 | –2.86 | 0.019 |
| Boxiness–$T_{30}$ | 0.8743 | 4.03 | <0.005 |
| Boxiness–EDT | 0.8933 | 4.44 | <0.006 |
| Boxiness–SEBR | –0.9221 | –5.33 | <0.005 |
| SQ–$T_{30}/T_{\mathrm{m}}$ | –0.8354 | –3.40 | 0.01 |
| Boxiness–$T_{30}/T_{\mathrm{m}}$ | 0.8722 | 3.99 | <0.006 |

* For seven rooms *t*-test calculations had 5 degrees of freedom (df).
† for definitions see text.

and the small-room EDT bass ratio (SEBR), using one-third-octave values as well, was defined as

$$\mathrm{SEBR} = 10 \log\left[\frac{\mathrm{EDT}(80\ \mathrm{Hz}) + \mathrm{EDT}(100\ \mathrm{Hz})}{\mathrm{EDT}(250\ \mathrm{Hz}) + \mathrm{EDT}(315\ \mathrm{Hz})}\right] \quad \mathrm{dB}. \tag{11}$$

Both quantities show better correlation in logarithmic form.

It is likely that the embedded error in the SEBR is rather high, as it uses a 10-dB slope to estimate the EDT at low frequencies. Nevertheless, its consistency between the three means and the similar trends of SBR, which utilizes the more stable $T_{30}$, give some confidence. The normal bass ratio, which shows poor correlation, is computed using one-octave band values and is more precise.

The last new quantity introduced here is the low–high ratio (LHR), which was optimized to give a higher correlation with the boominess ratings,

$$\mathrm{LHR} = 10 \log\left[\frac{T_{30}(50\ \mathrm{Hz}) + T_{30}(63\ \mathrm{Hz})}{T_{30}(3150\ \mathrm{Hz}) + T_{30}(4000\ \mathrm{Hz})}\right] \quad \mathrm{dB}. \tag{12}$$

A summary of all the parameters for all seven rooms is given in Table 9. A general remark is in order prior to any far-reaching conclusions. All new and old acoustical quantities introduced here and their respective correlations with the rated parameters are not necessarily related linearly in reality, as may be implied by the extensive use of the correlation concept. Most likely, they are not. The correlations here merely show that there is a strong relation, at least in the range of the values inspected. Bearing all that in mind, we proceed to examine the strong correlations, their validity, and any possible implications.

The music and speech ratings show different correlation patterns, and so their average combinations show composite correlations, depending on the weights of the partial ratings.

### 5.5 Sound Quality (SQ)

The relation between $T_{30}$ and the final SQ ranks is shown in Fig. 5. Error bars are added around the music points, which designate the minimum and maximum values recorded. The minima and maxima are relatively compressed due to the statistical manipulation described in the preceding.

Speech ratings show a preference of the lowest RTs available (talk studio), where the library is the only room that is preferred despite its higher mean $T_{30}$. The talk studio has already a very dry mean $T_{30}$ of 0.12 s. Does the same trend continue outside the range, that is, would the highest SQ be achieved in an anechoic chamber? One may speculate that the oppressive, dead nature of the anechoic room and the subsequent recording would not be preferable.

In the music ratings $T_{30}$ shows a different trend. Although no specific function is fitted to the data (they are interpolated for clarification only), it is clearly seen from Fig. 5 that SQ peaks at the narrow $T_{30}$ range of between 0.3 and 0.5 s and quickly drops above and, more slowly, below that range.

The library presents an interesting case, as its $T_{30}$ is not a single defining parameter. Looking at the aggregate performance of the library, its high SQ rating may be ascribed to three things. First, its relatively high volume combined with the rather dry acoustics for midrange and treble provides an unobtrusive environment. Second, the interior design of the room and the multitude of furniture and books increases the amount of scattering in the room significantly, compared to other rooms in the experiment. Third, it is a room that was not designed acoustically, definitely not for the purpose used here. It can thus be perceived as a more "natural" environment than a highly designed and artificial environment such as a studio or a listening room. In that sense, the meeting room performance, especially in the music ratings, can be interpreted as related to a somewhat more natural sounding (yet very little scattering). Another high correlation between SQ and SBR is illustrated in Fig. 6.

### 5.6 Boxiness

Boxiness can generally be considered an unwanted characteristic picked by listeners (see Fig. 7). The term boxiness was used in the first place as an indirect measure of coloration. However, the high correlation shown between boxiness and RT and EDT (Fig. 8), especially apparent in the speech ratings, cast a doubt over the connec-

Table 9. Room acoustical data.

| Parameter* | Control Room | Talk Studio | Hear. Prot. Room | IEC Room | Lecture Room | Library | Meeting Room |
|---|---|---|---|---|---|---|---|
| $V$ (m$^3$) | 105 | 82 | 27 | 97 | 186 | 186 | 85 |
| $T_{30}$ (s) | 0.195 | 0.125 | 0.825 | 0.358 | 0.526 | 0.283 | 0.495 |
| $T_{20}$ (s) | 0.200 | 0.125 | 0.825 | 0.364 | 0.534 | 0.282 | 0.494 |
| EDT (s) | 0.183 | 0.120 | 0.742 | 0.329 | 0.440 | 0.263 | 0.376 |
| $f_s$ (Hz) | 87 | 79 | 350 | 122 | 107 | 78 | 153 |
| BR(63, 125 Hz) | 1.173 | 1.36 | 1.175 | 2.042 | 2.118 | 3.449 | 2.658 |
| BR(125, 250 Hz) | 1.012 | 0.945 | 1.434 | 1.047 | 1.119 | 1.407 | 1.273 |
| SBR (dB) | 1.57 | 3.16 | −2.23 | 4.40 | 3.32 | 5.39 | 4.16 |
| SEBR (dB) | 1.55 | 3.34 | −1.30 | 2.76 | 2.03 | 4.10 | 1.68 |
| $T_m$ (s) | 0.2541 | 0.234 | 0.1616 | 0.2475 | 0.3075 | 0.3075 | 0.2368 |
| $T_{30}/T_m$ | 0.7699 | 0.5342 | 5.1111 | 1.4483 | 1.7111 | 0.9221 | 2.0908 |
| LHR (dB) | 2.09 | 2.66 | 0.11 | 5.08 | 6.55 | 6.51 | 4.88 |

* $V$—approximate volume; mean $T_{30}$, $T_{20}$, and EDT between 200 and 4000 Hz; $f_s$—approximate Schröder frequency; two alternative BRs using different octave bands in numerators. See Section 5.4 for other definitions.

tion to coloration or, more precisely, what subjects actually understood by a boxy sounding sample. In all cases boxiness had a strong correlation to the newly used small-room EDT bass ratio (SEBR), which relates bass to midrange frequencies (see Fig. 9). In large halls a similar bass ratio is associated with warmth and brilliance of sound. Do these qualities have any association with the suggested SBR and SEBR? It cannot be inferred from the

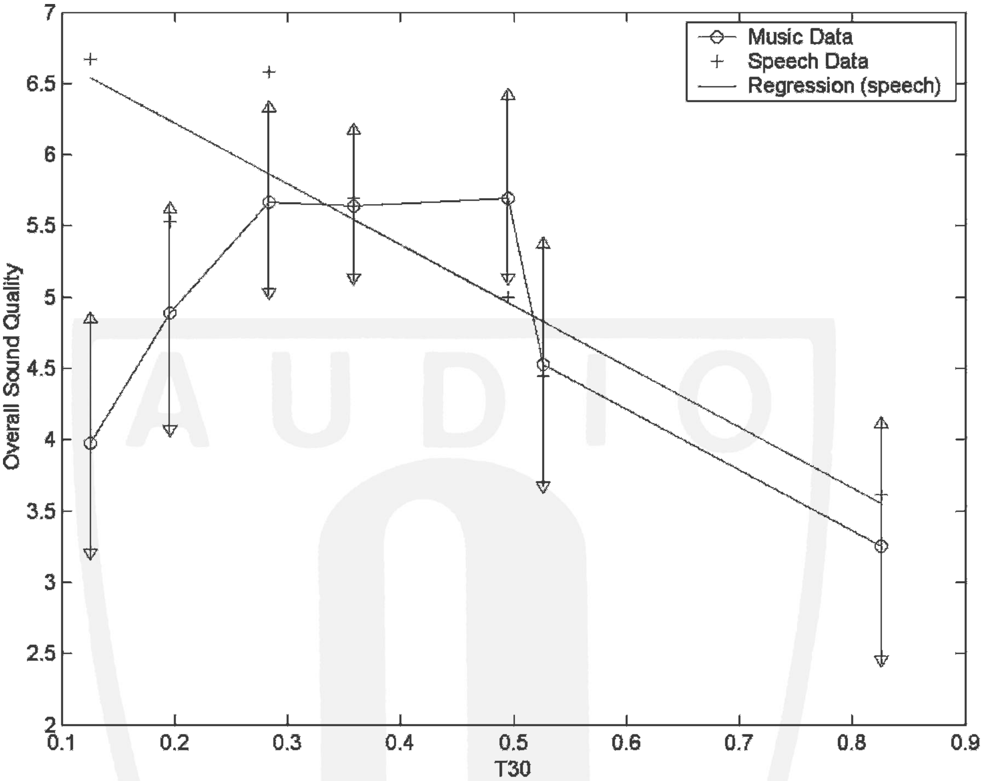

Fig. 5. Mean data for sound quality versus $T_{30}$. Regression line is for speech data. Music data (point error bars) are interpolated linearly. Regression yields $r^2 = 0.835$.

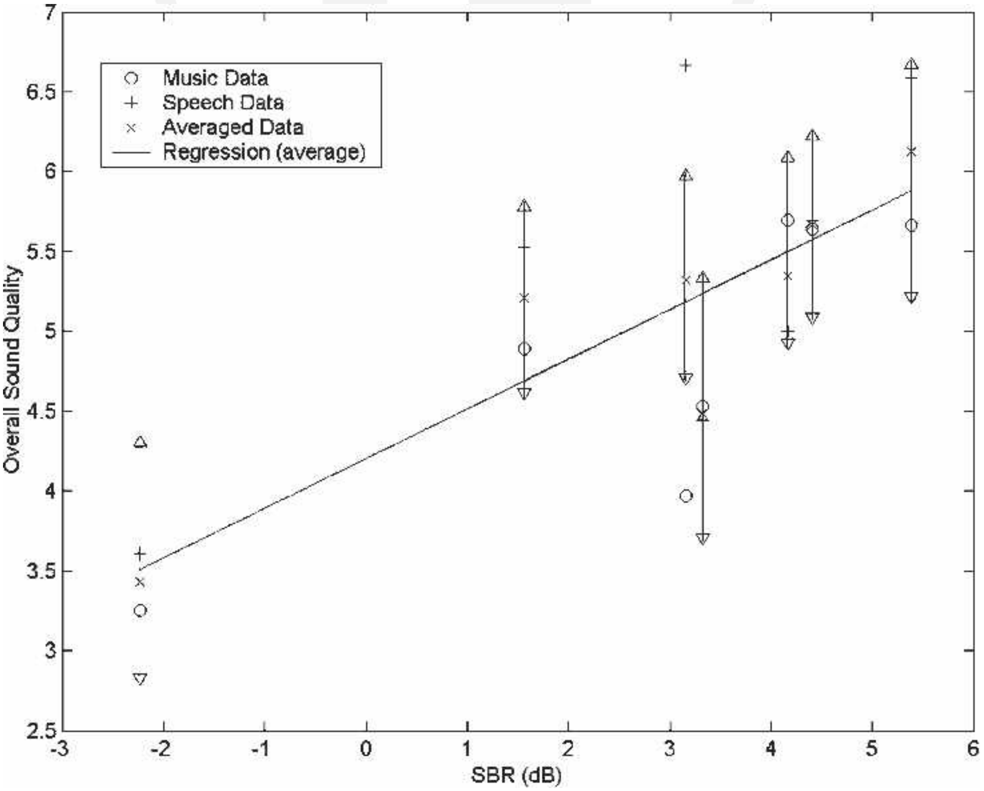

Fig. 6. Mean data for sound quality versus SBR. Regression line with point error bars is for averaged data. Regression yields $r^2 = 0.796$.

available data. The poor correlation of BR with any rating might imply some different meaning.

### 5.7 Boominess

Boominess could not be well correlated to any RT-derived quantity tested other than the newly introduced LHR (see Fig. 10). Only one room—the hearing-protector testing room—shifted the data from a monotonic functional behavior in the speech ratings, which in turn affected the composite score. This room has a different RT curve, which does not have a steep rise in the bass, but has a hump in the midrange (see Fig. 1). It is possible that in

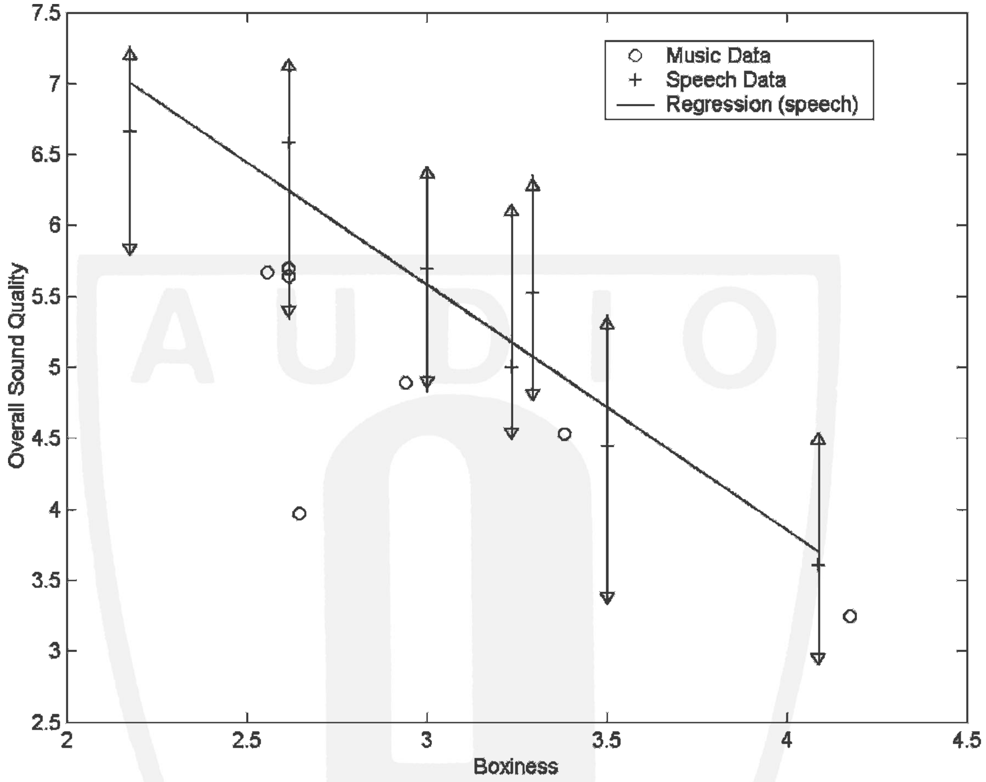

Fig. 7. Mean data for sound quality versus boxiness. Regression line with point error bars is for speech data. Regression yields $r^2 = 0.936$.

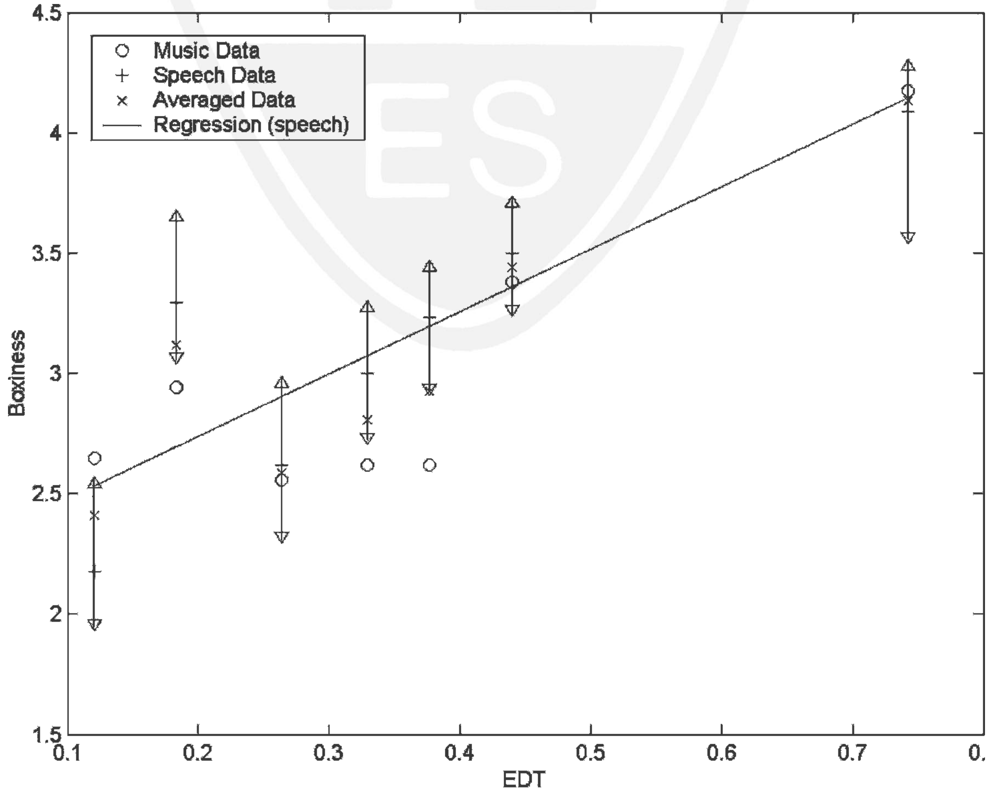

Fig. 8. Mean data for boxiness versus EDT. Regression line with point error bars is for speech data. Regression yields $r^2 = 0.936$.

the speech ratings some masking, confusing, or missing element was presented to the subjects, which made ratings more random, averaging at the scale center of 4. However, the entire boominess rating showed very small variance and was quite centered compared to the other two ratings, in addition to its lower reliability. Hence its accuracy is probably not high. Nevertheless, accepting the LHR as a meaningful quantification of boominess, one may relate it to some tonal balance measure—low bass versus treble—which can be perceived through the RT accentuation. See-

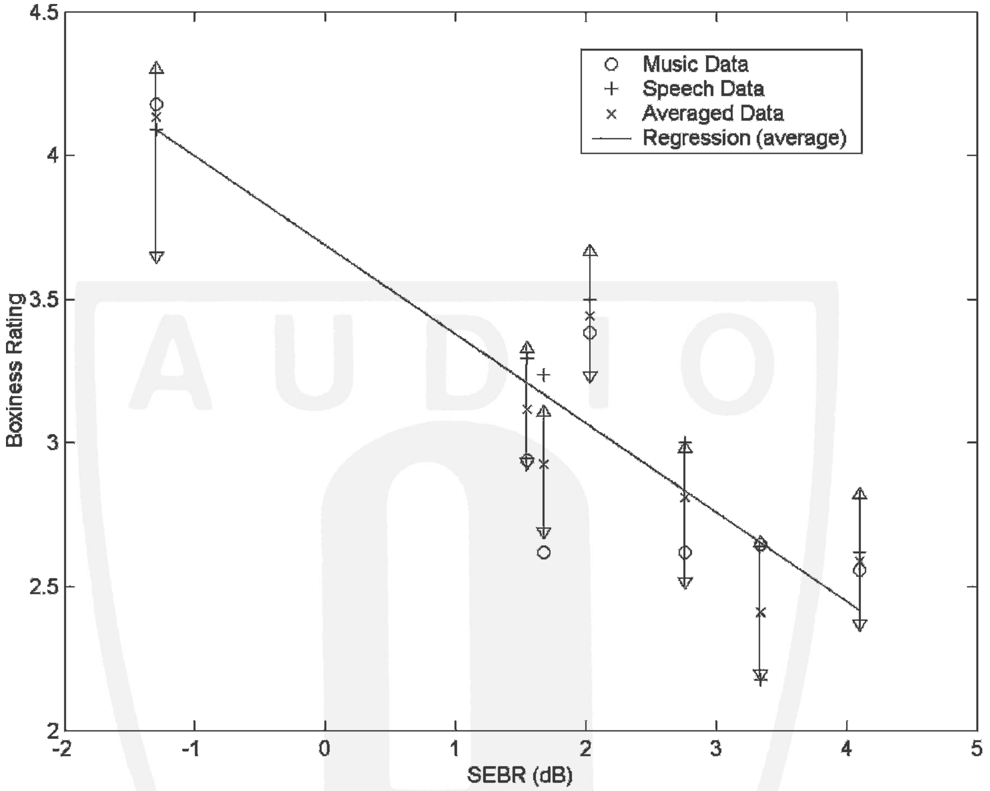

Fig. 9. Mean data for boxiness versus SEBR. Regression line with point error bars is for averaged data. Regression yields $r^2 = 0.85$.

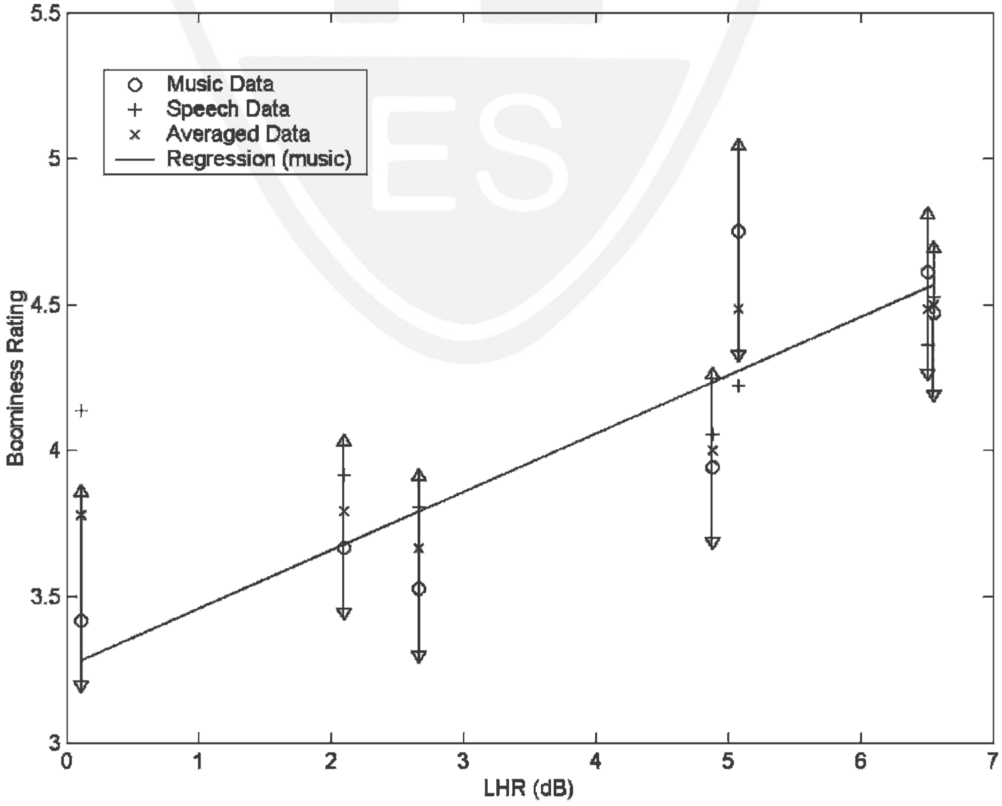

Fig. 10. Mean data for boominess versus LHR. Regression line with point error bars is for music data. Regression yields $r^2 = 0.774$.

ing that there is no clear functional dependence between the boominess and the SQ of the rooms (inspected separately for nonlinear dependence as well), this measure is of limited use. The lack of any correlation between SQ and boominess can be interpreted in a few ways: the subjects' difficulty in understanding the boominess concept, acceptable tolerance for a wide range of boominess available in rooms, inherently inadequate samples for boominess rating, or small objective variation between rooms. It may well be that the term boominess was ill-chosen to describe a bassy sensation, and its subsequent correlation with LHR is an alternative definition resorted to by the raters.

### 5.8 Multiple Regressions

The last step taken was calculating various multiple linear regressions between SQ, boxiness, and boominess and more than one acoustical variable at once. The results for speech and music are, once again, markedly different. Multiple regression between the speech SQ rating as the dependent variable and SEBR and LHR as two dependent variables showed a surprisingly high goodness of fit, with $r^2 = 0.9815$. It suggests that the sound quality for speech programs can be modeled based only on these two parameters, which are both derived from the RT of the rooms,

$$\begin{aligned}\mathrm{SQ} = b_1 \cdot \mathrm{SEBR} + b_2 \cdot \mathrm{LHR} + b_3 = 0.84 \cdot \mathrm{SEBR}\\ - 0.27 \cdot \mathrm{LHR} + 4.73.\end{aligned} \tag{13}$$

Interpolation between the points defines a plane. As was said before with regard to $T_{30}$, it seems reasonable to assume that the plane curves and reaches an optimum just before hitting the axes. Taking into account the sampling errors and the measurement uncertainties in RT, the $r^2$ might even seem exaggerated. Using the same two independent variables, boxiness can be predicted as well, with $r^2 = 0.9498$,

$$\mathrm{Boxiness} = -0.46 \cdot \mathrm{SEBR} + 0.14 \cdot \mathrm{LHR} + 3.49. \tag{14}$$

The music program shows a more complicated trend. No less than five independent variables are needed to model the SQ ratings for music linearly, a very dubious fit for only seven points, which is therefore rejected. Multiple regression between the SQ ratings and boxiness and boominess shows the best fit for speech, with $r^2 = 0.9236$, and a considerably worse fit for music, with $r^2 = 0.7931$. Hence either there are hidden parameters that were not gathered in this experiment, or the idea of a linear model assumption does not hold. Most likely it is both combined. The $T_{30}$ example described (Fig. 5, for music) supports a more general nonlinear dependence with SQ.

### 5.9 Summary of Main Findings

The main findings of the experiment are as follows.

1) Ratings usually showed different preferences for the speech and musical programs in terms of the overall sound quality of the recordings.

2) Classical bass ratio definitions showed unsatisfactory correlations with all ratings, and new quantities had to be sought.

3) Speech sound quality was generally rated higher for rooms with lower mean RT or EDT (200–4000 Hz).

4) A seemingly complete modeling of the speech SQ used two new parameters only, small-room EDT bass ratio (SEBR)—a variation on the classical bass ratio—and low–high ratio (LHR). The latter is also derived from the RT of the room, yet it describes the difference between the high-treble and the low-bass RTs in the rooms.

5) Music SQ ratings peaked at higher mean RT (and EDT) between 0.3 and 0.5 s. It also showed the highest correlation with another new parameter, the small-room bass ratio (SBR). However, the music SQ could not be modeled using only the parameters reviewed and rated, and it is likely that there is at least one hidden parameter, which not measured, that is instrumental to understanding the perceived SQ for music.

6) When music and speech are combined to one total SQ rating, it displays high correlations with both SBR and SEBR.

7) Boxiness showed less dependence on the program material. It inversely correlated well with SEBR in all cases and rather well with $T_{30}$ and EDT.

8) Boominess could only be correlated, to some degree, with LHR, especially in the music rating.

9) The mean RT, as recommended by the listening-room standards using the room volume, showed also some high correlations with SQ and, inversely, with boxiness.

Furthermore there are findings regarding specific rooms:

1) The library was the best room out of the seven, combining all ratings. It is rather reverberant at very low frequencies, but not so at midrange and treble bands. In speech ratings it scored second only to the talk studio, which has very dry acoustics at all frequencies.

2) The poorest two rooms are the lecture room and, even worse, the hearing-protector testing room. Both have high mean RT. The RT curve of the latter is irregular in shape, emphasizing the midrange frequencies over the rest. The lecture room has an almost unrestrained RT at very low frequencies because of the many axial modes in the room, which encounter little damping from the bare brick walls.

## 6 CONCLUSIONS

The three new quantities SBR, SEBR, and LHR show the strongest relationship to the SQ, boxiness, and boominess in the small rooms measured. Whether they are valid measures of other small rooms with different program material and perhaps different loudspeakers used, is a subject for further research. It is plausible that their functional form will be maintained, but perhaps with shifted frequency bands, which fit best the specific data.

However, if we choose to accept the specific frequency dependence exemplified in this experiment, it shows increased subject sensitivity to very low frequency bands, ranging from 50 to 100 Hz. It is unclear why the SBR give such different results in correlations and performance compared to the SEBR.

The ratings of boominess and boxiness did not fulfill the original intention. Boxiness ratings showed high dependence in the RT of the rooms, and thus it cannot be said whether they are associated with coloration in the room without further calculations of the room acoustics. Boominess might have been affected by the freedom to vary the presentation level, which changes the bass loudness perception nonlinearly. There is doubt as to how well the subjects understood both terms (logical error).

As for future research avenues, a simple survey can be made by using the existing measured RT curves of various rooms and calculating these parameters. They can then be compared with the known satisfaction from their performance. All in all, a more comprehensive questionnaire has to be presented to subjects, in which many more terms are used to describe the room acoustics quality. A research similar to the one done by Gade [13], [14] about musicians' conditions, can then account for all the factors that define the overall sound quality in small rooms. For example, an important subjective parameter, which was left out here, is timbre.

It is arguable whether the testing method in this listening test was the most effective one. For instance, paired-comparison methodology would probably give more reliable results, especially for the less reliable and clear boxiness and boominess concepts. However, the results from the testing method chosen could usually be interpreted in a sensible way.

The initial experimental design was an attempt to deal with a few known issues in small-room acoustics. Not all were treated and some are left unanswered:

1) New parameters were introduced in order to relate the preferred sound quality to the objective acoustical data of the rooms. They seem to depict the subjective acoustic quality of rooms better than existing parameters, however partially. Why they do so with greater success is not entirely understood.

2) The nonflat RT curve issue can be addressed to some extent. Repeating the introductory question: is a longer (on average) yet flat RT curve preferable over a short mean RT with a longer reverberant bass? The room survey is not comprehensive enough to review more than a few combinations. However, it seems that the answer is negative. It comes from examination and inference from the opposite case: a flat short RT curve (talk studio and control room) versus a longer nonflat RT curve (library, IEC listening room, and even the meeting room). Subjects preferred the more reverberant bass of the latter group over the drier former group in the music ratings. In speech ratings, the results were mixed, but the library and talk studio are comparably good. The combined ratings show a clear preference of the more reverberant room group.

## 7 ACKNOWLEDGMENT

The authors would like to thank all the people who made this project possible for their kind help: Jørgen Rasmussen and Tom Petersen for the endless technical support in the acoustics laboratory; Henrik Haslev, Brüel & Kjær, for kindly loaning us some essential equipment; Lars C. Bruun, DR Radio, for letting us measure in the radio's studios; Klaus Kaae Andersen for statistical guidance; Brent Kirkwood for numerous advice and general help; Yehuda Spira for the room sketches; Jan Voetmann, Delta A/S, for providing the initial topic, additional direction of the project, and useful comments; Benjamin Fenech and James Black for more help.

## 8 REFERENCES

[1] H. Kuttruff, "Sound Fields in Small Rooms," in *Proc. AES 15th Int. Conf.* (Copenhagen, 1998 Oct. 31–Nov. 2), pp. 11–15.

[2] M. Vorländer, "Objective Characterization of Sound Fields in Small Rooms," in *Proc. AES 15th Int. Conf.* (Copenhagen, 1998 Oct. 31–Nov. 2), pp. 16–23.

[3] ISO 3382:1997(E), "Acoustics—Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters," International Standards Organization, Geneva, Switzerland (1997).

[4] ISO R-1996–1971(E), "Acoustics—Assessment of Noise with Respect to Community Response," App. Y, International Standard Organization, Geneva, Switzerland (1971).

[5] ANSI S12.2-1995, "Criteria for Evaluating Room Noise," American National Standards Institute, New York (1995).

[6] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Evaluation of Artificial Heads in Listening Tests," *J. Audio Eng. Soc.,* vol. 47, pp. 83–100 (1999 Mar.).

[7] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer Characteristics of Headphones Measured on Human Ears," *J. Audio Eng. Soc.,* vol. 43, 203–217 (1995 Apr.).

[8] D. Pralong and S. Carlile, "The Role of Individualized Headphone Calibration for the Generation of High Fidelity Virtual Auditory Space," *J. Acoust. Soc. Am.,* vol. 100, pp. 3785–3793 (1996 Dec.).

[9] J. P. Guilford, *Psychometric Methods* (McGraw-Hill, New York, 1954), pp. 278–297, 394–397.

[10] ITU-R BS.1116-1, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (1994–1997).

[11] Leo L. Beranek, *Music, Acoustics & Architecture* (Wiley, New York, 1962), pp. 433–436.

[12] EBU Tech. 3276, "Listening Conditions for the Assessment of Sound Program Material: Monophonic and Two-Channel Stereophonic," 2nd European Broadcasting Union (1998 May).

[13] A. C. Gade, "Investigations of Musicians' Room Acoustic Conditions in Concert Halls. Part I: Methods and Laboratory Experiments," *Acustica,* vol. 69, pp. 193–203 (1989).

[14] A. C. Gade, "Investigations of Musicians' Room Acoustic Conditions in Concert Halls. Part II: Field Experiments and Synthesis of Results," *Acustica,* vol. 69, pp. 249–262 (1989).

## THE AUTHORS

A. Weisser

J. H. Rindel

Adam Weisser was born in 1978 in Haifa, Israel. He received a B.A. degree in physics from Technion, Haifa, in 2000. Later he received a master's degree in engineering acoustics from the Technical University of Denmark, Lyngby, in 2004.

He currently works for Oticon A/S in Smørum, Denmark.

●

Jens Holger Rindel was born 1947 in Copenhagen, Denmark. He received an M.Sc. degree in civil engineering in 1971 and a Ph.D. degree in acoustics in 1977, both from the Technical University of Denmark in Lyngby.

He is an associate professor, and since 1990 he has been professor in acoustics at the Technical University of Denmark. During leaves from the university he worked as a senior researcher at the Norwegian Building Research Institute in Oslo and as a visiting professor at the University of Sydney, Australia, the Nihon University, Japan, and the Acoustic Research Centre, University of Auckland, New Zealand.

Dr. Rindel has been active in room acoustic research, first as a project leader for the development of a room acoustic computer model (Odeon software), which is today used worldwide by major consulting companies for acoustic design, and then with many universities for education and research.

He is a fellow of the Acoustical Society of America and the Institute of Acoustics, UK.