Sound Quality in Small Room

Adam Weisser

March 8, 2005

Abstract

In this project three novel experiments in small room acoustics are presented. The experiments employ samples from a specially recorded binaural library of seven different small rooms of both good and bad acoustics.

The first experiment relates subjective ratings of overall sound quality from the recordings to the reverberation time in the room and to other related objective parameters. It also examines ratings of room "boxiness" and "boominess". The ratings are inspected separately for speech, music and both combined.

The other experiments examine the capability of subjects to distinguish between recordings from different positions in one room to those made in another room. In spite of noticeable timbral differences in all recordings, it is shown that subjects are able to make the distinction in many cases. It is also shown that the defining parameter in this process, is most likely a quantity derived from the reverberation time difference between the two rooms.

* * *

I dette projekt presenteres tre nye eksperimenter i små-rum-akustik. Eksperimenterne anvender prøver fra et specialoptaget binarual samling af syv forskellige små rum af både gode og dårlige akustik.

Det første eksperiment forholder sig til subjektive vurderinger af den generelle lydkvalitet fra optagelserne til efterklangstiderne i rummet og til andre relateret objektive parametre. Derudover efterforsker det også vurderinger af rum-boxiness og rum-boominess. Vurderingerne undersøges adskilt med hensyn til tale, musik og begge kombineret.

De andre eksperimenter eksaminerer evnen af forsøgpersoner til at skelne mellem optagelser fra forskellige positioner i et rum og dem udført i et andet rum. Til trods for bemærkelseværdige klangfarvet forskelle i alle optagelser viser det sig at forsøgsepersoner har evnen til at kunne skelne i mange tilfælde. Det er også vist at det afgørende parameter i denne process er mest sandsynlig en kvantitet udledet fra efterklangstidsforskellen mellem de to rum.

Acknowledgments

This project is the final master thesis of my degree in the Acoustic Laboratory in Lyngby, Denmark. It owes its completion to many people, to whom I am indebted for their help and support:
Jørgen Rasmussen and **Tom Petersen** for the endless technical support in the acoustical laboratory.
Henrik Haslev, Brüel and 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 the statistical guidance.
Brent Kirkwood, for numerous advices and general help.
Yehuda Spira, for the nice room drawings.
Michael Yang, for the help in Danish.
Torben Poulsen, for some good advices in psychoacoustics and for making the initial connexion with Delta.
Anders Christian Gade, for preliminary statistical help.
Stephan Ewert and **Oliver Fobel,** for a little help in Matlab.
Hiroshi Onaga, for extra volunteering for the test.
Bruno Fazenda, for some stirring discussions.
Torben Peterson, Delta A/S, for connecting me with Jan Voetmann.
Jens Holger Rindel, my supervisor, for the patient and always helpful guidance of this project.
Jan Voetmann, Delta A/S, for providing the initial topic and additional direction of the project.
Oticon A/S and **the Danish government,** for sponsoring my stay in Denmark during the degree.
Benjamin Fenech, James Black and **Claudiu Pop,** for the casual help and the many laughs.
Søren Bech, Bang & Olefsen; Flemming Christensen, Aalborg University; Russ Berger, Russ Berger Design Group; and **Gavin Haverstick, Auralex Acoustics, Inc.,** for promptly replying my inquiries.
Plus, a special shout out to my family and all other friends at home and abroad, who continuously supported me throughout the degree.

Contents

Chapter 1

Introduction

The following topics are covered in the introduction:

- Brief evolutionary background of the project with a description of the questions it intended to address through research
- Experimental design background
- Summary of important uses for small room as acoustical environments
- A critical review of a similar research
- Project overview

1.1 Research Background

The present project evolved from discussions between Jan Voetmann, Jens Holger Rindel and myself, which yielded a draft for an experiment based on practical needs of the acoustician facing a small room design and, also, on the existing publications done in the small room acoustics field. The main aim of the experiment is to quantify preferences of listeners of various small room acoustics and relating them to the objective acoustical properties of the rooms. Although very general and rather rudimentary in its goal, such an experiment poses many immediate problems and thus entails careful consideration of all elements concerned.

The experiment was designed in attempt to answer the two broad questions:

- Why does the sound quality of some small rooms is superior to others?
- How well do the existing acoustical measures (and design standards for that matter) describe, qualify and quantify the sound quality in small rooms?

These same questions in large rooms and halls are much more researched. However, there are many essential differences between the acoustical behaviors of small and large rooms and thus, immediate use of the same concepts used in the design of large rooms cannot be made. Small rooms, despite being in daily use, are less attractive to the acoustician, apart from specific cases such as studios, listening rooms and control rooms. Even for these types of rooms there is no agreement between acousticians what exactly is best and most appropriate and how much freedom there is in setting different parameters within the small rooms. Examples for that can be shown in the listening room standards, which show the acceptable wide tolerance in reverberation times of rooms at low frequencies.

Two recurring problems were brought up that have special interest:

- Long reverberation can mask some coloration at low frequencies, but it is unclear what is their exact interrelation.
- Does a balanced, flat curve of longer midrange mean reverberation time is preferable to a shorter mean RT, but longer at the bass bands? There are some hints and intuitions, but no well-established answer.

1.1.1 Experimental Design

The approach agreed upon in order to tackle as many topics as possible, is to survey a selection of small rooms, which vary in shape, use, construction and hence in their acoustical properties. Sound samples would be played back in the rooms and recorded binaurally using a dummy head. In addition, the acoustical parameters of the rooms would be measured. The recordings are to be introduced later to test subjects in a series of psychoacoustical evaluations (Experiment A), rating their overall preference (later it will be referred to as overall sound quality) and capability to hear different characteristics of the recordings (later - boxiness and boominess), which necessarily stem from the various room acoustics. The final results will be statistically analyzed in order to find possible correlations between the subjective evaluations and the acoustical properties of the rooms.

Special attention must be given to the following points:

1. Received acoustically "good" rooms, as well as acoustically "bad" rooms must all be measured. That ensures a wide array of results, which may be easier to correlate later. Also, it gives the subjects the unbiased option to grade bad rooms and to establish their own ranking rationale.
2. The use of binaural recordings is problematic. Binaural recording technique is under constant research and the employment of dummy head must be done carefully, taking into account the various side effects of that method. Similar-in-effectiveness alternatives to binaural recordings are either auralized simulations or measurements in real rooms (or one room with variable dimensions and treatment). Auralization is not yet in a state comparable to recordings in their degree of realism. If high-quality simulations are possible, they would surely demand more working time to construct than binaural recordings. The in-situ measuring option is appealing - first hand measurement - but introduces logistical problems of gathering all the subjects in small rooms, and in many cases relying on short term acoustical memory of subjects if pair comparisons are to be made.
3. Precise and focused phrasing of the subjective evaluation questionnaire must be pursued. Failure to do that may result in either biased or useless results.

The two other experiments (B and C) were not a part of the original design. The choice to perform them was largely made because they were readily available. The recorded library that was formed for the preference test is suitable for constructions of many tests and the options followed here are only a few of them. Experiments B and C were formed to test the question: Do listeners have the ability to distinguish between a room and itself? Using different recordings from the same room only at different positions a matching test was constructed. Subjects had to match two recordings made in the same room having recordings from other room as as alternative in a forced-choice test. Furthermore, it was asked, does a room have a unique imprint on the sounds conveyed in it?

1.2 Special Types of Small Rooms

As a brief background for the importance of small rooms acoustics, the following describes a few main types. Among the numerous types and uses of small rooms, some demand careful acoustical consideration if there is professional work to be done there, which involves high quality sound reproduction and recording and scrutiny. In addition, acoustical design standards evolved from the need to be able to replicate the same listening conditions in different rooms.

1.2.1 Talk Studios

Talk studios are indispensable for almost any radio station. They are different than other studios, like music studios, by the requirement for very short reverberation time and natural sounding timbre of the people talking, i.e. without coloration. These conditions ensure high speech intelligibility and a pleasant, unoppressive listening experience, as far as the room acoustics is concerned. Two examples of talk studio design are given in depth in [1].

Another important application for talk studios is for film dubbing and automatic dialogue replacement (ADR), a post-production dubbing used in many contemporary movies. Talk studios provide dead acoustics, which ensures both high speech intelligibility and relative ease in later adding artificial reverberation to simulate the environments, which are shown in the pictures.

1.2.2 Control Rooms

Control rooms are the rooms where the mixing and production of music recordings are done. The sound technician and producer, sitting in the control room, have to hear with absolute high quality, high fidelity the recordings that are being done in the adjacent studio. For that purpose the equipment used has to be state of the art, free of distortion, noise, etc. to the highest attainable degree. Nonetheless, there is so much that the equipment can offer without complementary adequate room acoustics. An example of how the control acoustics participated in the mixing process is vividly described by Børja [2].

An ongoing dispute has been taking place in the last decades of how the control rooms should be designed in order for the final product to be as good as possible. The main issues are the reverberation time in the different parts of the rooms, reflections from all surfaces, including the mixing console, loudspeaker placement, coloration and accurate spatial image reproduction. The opinions and philosophies expressed and discusses over the years have changed radically, from a completely dead, all-absorbing room to more live rooms. Overview of these different approached and additional references are to be found, for instance, in Newell's article [3].

Of special interest is Walker's design of control rooms, [4] and [5], which employs principles of geometrical acoustics to create a reflection free zone for the first $15ms$ after the direct sound. It is related directly to the conclusions from Bech's research (see section 2.4.2). The idea is to maintain the room reverberant ambiance, restricting the use of absorption, while still avoiding coloration from the early reflections.

1.2.3 Listening Rooms

Standardized listening rooms are used for listening tests, where non-biasing, repeatable, acoustical and visual optimal conditions can be provided. Such tests include high-fidelity equipment evaluation, small impairment detection in digitally compressed audio media and more. Although similar to domestic listening rooms (or living rooms), it is unlikely that such acoustical conditions for stereo reproduction are met in most houses, but at the most dedicated aficionados.

The standards ITU-R Recommendation BS.1116-1, EBU Tech 3276 and Tech 3276E (for multi-channel reproduction), describe rather similarly the requisites for listening rooms [6], [7] and [8]. The ITU standard relates mostly to standardized tests of small impairments detection type of listening tests, where the listening is only a part of an array of demands to ensure unbiased and results, as far as the general conditions are concerned. The EBU standard relates to general critical listening conditions demanded in order to have a set of standard reference listening conditions in all rooms. It also mentions some control room design guidelines. The room acoustics specifications are rather similar in both standards and they prescribe the room dimensions and proportions, reverberation time, early and late reflections, background noise level and the so-called operational room response curve.

1.3 Spikofski's Experiment

This is a brief critical review of a similar EBU (European Broadcasting Union) commissioned research that was done a few years ago. It dispels an initial concern that the present research is an unnecessary repetition of the same experiment.

The paper written by Spikofski [9] describes a remarkably similar experiment to the one originally intended to be performed in the present project. The similarities are: binaural recordings of several identical programs in different small rooms, subjective scaling of listener's preferences and attempt to correlate them to objective parameters of the room acoustics. The major and immediate differences are: focus only on control rooms with high standards (recorded in state of the art radio stations across Europe), use of the control room monitors for stereophonic reproduction, closer inspection of background noise, tonal imbalance and image quality. Also, an elaborate head tracking system was utilized, to perfect the binaural recording reproduction. The main interest of the research is a review of the existing listening room standards [6], [7] and [8].

This research has one important conclusion in regard to the so-called neutrality of the control rooms. The subjective tests show dissimilarities between the rooms, suggesting that the tolerances specified in the listening room standards are too broad.

In addition, some parts in that paper seem rather questionable and are presented without discussion. For instance, the separate definitions of coloration and tonal balances and their succeeding mix-up synonymously; the choice used to correlate subjective coloration perception with an objective figure, squaring all the deviations from a "loudspeaker/room response curve". Upon failure to correlate the two using this method, the author deems the mission unattainable to compute in simple means; the choice of the particular scaling method with reference to another room and later giving an equal weighting to each subjective parameter in the overall grading of the recordings.

1.4 Project Overview

The following chapters in this report are constructed in the following manner:

- Chapter 2 - Theoretical background for the small room acoustics filed and a short account of available research
- Chapter 3 - Description of the experimental methods that were used in the research, notably the binaural recording technique and the psychometric testing methods
- Chapter 4 - Account of the reverberation time measurements and sample library recordings in the surveyed rooms
- Chapter 5 - Listening test construction and task description
- Chapter 6 - Results and analysis of Experiment A
- Chapter 7 - Results and analysis of Experiments B and C
- Chapter 8 - Discussion and conclusions

Chapter 2

Theoretical and Experimental Background

The theory behind small room acoustics is as vast as and no different than that of any other room acoustics. The basic theory is briefly covered here with emphasis both on relevant concepts for later use in the project and also on the main differences between large and small room acoustics. All that is intertwined with experimental findings available in the literature. The following subjects are covered:

- a brief overview of the approaches to acoustical analysis of enclosures
- the wave equation
- statistical acoustics
- small room phenomena and
- relevant subjective measures of small enclosures for subjective characterization.

2.1 Overview - Approaches to Acoustical Analysis of Enclosures

The acoustical analysis of enclosures can be performed in three ways, which treat the sound propagation differently.

The first analytic approach involves solving the sound wave equation contained within the enclosure's boundaries. This approach is more exact and makes less approximations than the other methods. However, there are two serious problems with extensive use of this method. First, the wave equation is solved in 3-dimensions and the solution becomes immensely complicated and sensitive to small errors as the frequency increases, as is shown below. Second, the wave equation can be solved analytically only for a very limited types of enclosures. The analytic solution for very simple room shapes is available. But not for most other practical structures and certainly not in an analytic, instructive form that can be interpreted intelligently later. Other structures have to be solved numerically with methods such as Finite Element Method (FEM), Boundary Element Method (BEM) or time domain analysis.

At high frequencies, when the wave equation approach becomes too complicated, statistical acoustics can be used. It is based on certain approximations that greatly facilitate part of the problem solving, mostly for general room acoustics parameters.

Geometrical acoustics makes use of the concept of sound rays, similarly to light rays in geometrical optics. The sound propagates from the source and hits surfaces whereupon part of it is reflected, refracted and loses energy until it becomes insignificant. This description is made more accurate using directionality

patterns of the sources and frequency dependent absorption and scattering of the various surfaces. This geometrical approach may benefit by using some of the premises of statistical acoustics, which are generally correct for large rooms and high frequencies. The geometrical approach does not take into account wave phenomena such as diffraction. Using it in other situations leads to erroneous results where the wavelength and phase cannot be neglected, as is inherently implied from the use of sound rays propagation. Nonetheless, the use of the sound rays concept is especially suitable for computer-aided analysis, where these relatively simple principles provide adequate solutions to highly complicated rooms such as concert halls [10].

In reality, sound does propagate as waves and not as rays. Geometrical acoustics simplify situations where the exact wave phenomena are not necessary in order to correctly predict the behavior of the sound field, for example echo, focus and various other types of reflections.

2.2 The Wave Equation and Rooms Modes

Let us go over the fundamentals of the solution of the three-dimensional sound wave equation, through which all of the above would be more clearly explained. In the following overview some key concepts and formulas are presented. The following is based mainly on Jacobsen [11] and Kuttruff [12], where the reader can find a more rigorous and complete derivations, with relevant references.

Any enclosure is defined by its boundaries, characterized by their complex acoustical impedances, which in turn dictate the boundary conditions on the sound pressure and its gradient. Generally, the sound field is expressed by the Helmholtz wave equation:

$$\nabla^2 \hat{p} + k^2 \hat{p} = 0 \tag{2.1}$$

Where $\hat{p}$ is the complex pressure and k is the wave number, which is related to the frequency and wavelength in a particular way, depending on the exact situation.

When the room is lightly damped (referred to as a reverberation room) – the walls are either rigid or reflect most of the energy – the solution to (2.1) with these conditions is a set of eigenfunctions ("mode shapes"), each with its respective eigentone (a "natural frequency" of the room). The modes are in fact resonances of the room and they have a specific geometry within the room. At every such room mode there is a possibility for a standing wave to evolve, when the source frequency coincides with the natural frequencies of the room modes. A standing wave is formed when a wave is reflected from a boundary so that the nodes and the antinodes of the incident and reflected waves are in phase. A standing wave does not move in time, but only its amplitude changes in time.

An algebraic property of all the functions that solve (2.1) with its particular boundary conditions, is that they are orthogonal to each other. Hence the alternate name "normal modes" for the eigenfunctions in reverberation rooms. Thus, for any eigenfunction Ψ,

$$\frac{1}{V} \int \Psi_m \Psi_n dV = \delta_{mn} \tag{2.2}$$

Normalized and integrated over the entire volume of the room, V.

The complete sound pressure is obtained by summing up all the different eigenfunctions.

$$\hat{p}(x, y, z, t) = \sum_N A_N \Psi_N(x, y, z) e^{j\omega t} \tag{2.3}$$

Where (2.3) shows the different eigenfunctions summed up with different weight coefficients A_N and a with harmonic time dependence $e^{j\omega t}$.

An analytic solution to the wave equation (2.1) is available only for simple shapes such as boxes, spheres and cylinders. The box shape, in Cartesian coordinates, is of particular importance in rooms, being so commonly built in a box shape, and it has an instructive value. Assuming all walls are infinitely

hard, the pressure gradient is zero at the walls (i.e., the sound velocity normal to the wall is zero). The eigentones of such a room, f_N, are given by (2.4):

$$f_N = \frac{c}{2}\sqrt{\left(\frac{n_x}{l_x}\right)^2 + \left(\frac{n_y}{l_y}\right)^2 + \left(\frac{n_z}{l_z}\right)^2} \tag{2.4}$$

Where c is the speed of sound in the room, n_x, n_y and n_z are integer numbers (running from 0) corresponding respectively to the room axes; and the room are dimensions l_x, l_y and l_z. The respective mode shape functions would be of the form:

$$\Psi_N(x,y,z) = \Lambda_N cos\left(\frac{n_x \pi x}{l_x}\right) cos\left(\frac{n_y \pi y}{l_y}\right) cos\left(\frac{n_z \pi z}{l_z}\right) \tag{2.5}$$

The function depends on the location in the room as of the coordinates x, y and z. Λ_N is a normalization factor.

The eigenmodes are divided, generally, into three types. The axial modes are one-dimensional, for which two out of the indices n_L, n_W and n_H are equal to zero. Tangential modes are two-dimensional, where one index is equal to zero, and oblique modes, where non of the indices is zero. When all the indices are zero the fundamental mode is present and the sound pressure is independent of the position. Axial modes represent the simple standing waves case, where a mode exists between two parallel walls. In tangential modes the standing waves are built up using reflections from four walls. In oblique modes all 6 walls contribute to the mode buildup.

For instance, the axial mode algebraic set relating to the l_x dimension of the boxed-shaped room is:

$$\frac{c}{2l_x}, \frac{c}{l_x}, \frac{3c}{2l_x}, \frac{3c}{l_x}, \dots \tag{2.6}$$

In reality, since the walls have finite impedance, some energy is absorbed in the walls. If the absorption is not too high, orthogonality of the eigenfunctions can still be assumed. Every eigenmode would then be complex and the eigenmode function will have its characteristic width. This behavior is typical for any resonant system. The modal bandwidth, Δf, then describes the spectral width of the mode, which is related to the energy loss at the walls - damping. In practice it means that a certain mode of the room will be noticeable for source frequencies, which do not coincide exactly with the particular natural frequency. Putting it differently, all frequencies in the room are carried through the room modes, be it in their natural frequencies, or around it in the sidebands. The room modes are the only means of conveying sound energy in the room. The bandwidth of such a mode can then be approximated by

$$\Delta f \approx \frac{1}{2\pi\tau} \tag{2.7}$$

Where τ is the particular decay time of the mode, which depends on the wall impedance at that frequency (in higher frequencies there are also increasingly significant losses from the propagation in air).

A sound source placed in a room is coupled to a receiver through the room normal modes. As all the mode functions are algebraically orthogonal, (2.2), there is an individual coupling coefficient for each mode, which does not depend on other modes. Therefore, the source and receiver positions in a room are interchangeable. The receiver's response in any position ($\vec{r'} = x', y', z'$) can be expressed using the eigenmode functions. The source-receiver coupling coefficient is given by

$$C_i(\vec{r}, \vec{r'}) = \Psi_i(\vec{r})\Psi_i(\vec{r'}) \tag{2.8}$$

The overall sound pressure sum changes into:

$$p(\vec{r}, \vec{r'}) = \sum_{n=1}^{\infty} \frac{a_n C_n(\vec{r}, \vec{r'})}{k^2 - k_n^2 - ik_n/(\tau_n c)} \tag{2.9}$$

Where a_n is the relative amplitude of the n'th mode. The complex dependence can be seen in the denominator when the mode decay time is inserted τ_n into the equation.

It can be inferred from (2.8) and (2.9) that at mode nodal or anti-nodal points in the room, certain frequencies would be inaudible or infinitely loud at the receiver's position, if the sound is undamped (i.e., τ is infinitely long). In reality, since there are losses in every room, a particular frequency is never inaudible or infinitely loud. These losses are expressed through the finite τ in (2.9).

2.3 Statistical Room Acoustics

As the frequency increases, more and more modes come into play and contribute to the total pressure. Thus, in every room, there are three frequency ranges in terms of the modes. In low frequencies, below the first eigentone frequencies, sound is carried on the low sideband of the lowest mode. The farther the frequency is from that mode, the weaker it will sound, as the mode is hardly excited. Above that range some modes begin to show up in increasing density (in terms of number of modes per unit frequency). At this range the modal density is sparse and individual mode are likely to be dominant in the room at certain frequencies. The third range is reached, where there is a more uniform mode density and it is practically impossible to distinguish between contributions of individual modes, for they overlap each other. In this last frequency range a statistical approach is very useful in acoustics, where it is possible to make probabilistic predictions with little or even no knowledge about the room properties.

In a room with volume V the modal density can be approximated by

$$n(f) \approx \frac{4\pi V}{c^3} f^2 \tag{2.10}$$

Using the modal bandwidth (2.7) with the modal density (2.10), the modal overlap can be obtained. It measures the average number of modes, which are excited by a pure tone within their modal bandwidth. Experimental experience shows that large enough a modal overlap is three, in order to justify the use of the statistical approach.

A useful simple statistical model of lightly damped rooms assumes a perfectly diffuse sound field. It occurs where there are distant sources, which emit random noise in all directions. The resultant so-called diffuse field consists of sound waves in all directions, which are completely uncorrelated, and the field is then homogeneous and isotropic. It is also free from any interference effects due to its uncorrelated nature. This type of field is only theoretical and no real room has such properties. However, many situations can be approximated to having a perfectly diffuse sound field, facilitating the analysis while still maintaining accurate predictions.

One definition of special importance is that of the reverberation time (RT) in a room, T_{60}. It measures the time it takes for a steady sound to decay by $60dB$ when turned off in a room. In diffuse sound fields T_{60} can be expressed simply using Sabine's formula:

$$T_{60} = \frac{55.3V}{cA} \tag{2.11}$$

Where T_{60} is a function of the volume V, and the equivalent absorption area of the room, A, which is defined as:

$$A = \sum_i S_i \alpha_i \tag{2.12}$$

Where S_i and α_i are a surface area and its mean absorption coefficient, respectively. The absorption is frequency dependent and therefore, the RT is normally measured in octave or one-third octave bands.

A correction to Sabine's formula, which deals with highly absorbent rooms, is Eyrling's formula, which correctly predicts zero RT in anechoic chambers:

$$T_{60} = -\frac{55.3V}{cS\ln(1-\alpha_m)} \tag{2.13}$$

Where α_m is the mean absorption coefficient of the room. At low absorption the two formulas give the same results. Still, Sabine's formula is more recommended for use in general cases of various mixed rooms absorptions.

Both formulas can be corrected for the effect of the air absorption on the RT. The corrections are not given here.

Schröder's frequency is the frequency above which there is sufficient modal overlap, so that the statistical approach can be used in a certain enclosure.

$$f_s = 2000\sqrt{\frac{T_{60}}{V}} \tag{2.14}$$

In practice, decay measurements are prone to large errors if the entire $60dB$ decay is measured. Therefore, it is customary to measure a 10, 20, or 30 dB slope and extrapolate the rest, assuming a linear decay. For example, T_{20} is defined as:

$$T_{20} = 3 \cdot (t_{-25} - t_5) \tag{2.15}$$

Where t_x is the time when the sound has decayed by xdB after being turned off. Note that the measurement begins only after the initial $5dB$ decay, in order to have a linear decay curve without an offshoot.

2.4 Small Rooms

Much of the following discussion relies on references by Walker [13], Kuttruff [14] and Vorländer [15].

Rearranging (2.14), the following inequality is reached, which suggests a definition for the "acoustical size" of the room:

$$V < \left(\frac{2000}{f}\right)^2 \cdot T_{60} \tag{2.16}$$

Setting up a low frequency threshold from which statistical acoustics can be safely used, the volume of the room is defined, if the reverberation time is known. There is a direct relation between the reverberation time and the volume of the room. If the mode-dominated audible range is to be minimized, then either the RT should be decreased or the volume of the room increased. As there is normally a preferable RT for a specific use of the room and its volume is limited by practical constraints, a whole array of small room acoustics problems is created, simply because of the wider than optimal frequency range, which is dominated by the individual room modes (see next section). The inequality should not be taken as a clear-cut limit, but more as an indicator of the different frequency ranges, depending on the given RT's. An illustration of threshold frequencies for various RT's is shown in Figure 2.1. The graphs show how at very large room volumes, the mode-dominated range rapidly shrinks and converges for all RT's.

As the above inequality is not definite, a more straightforward alternative for small room definition may be used and that is simply according to its volume. Typically rooms up to $100m^3$ in volume are considered to be small, although rooms up to $200m^3$ may also fall into that category.

The small room acoustics can be viewed either from a frequency domain or a time domain point of view. Each one has its share of possible problems, which are sometimes complementary. Where small rooms have myriad purposes for everyday use, the existence of these problems may not be noticed at all in many cases, or sometimes accepted unconsciously. Professional use of rooms for sound recording and reproduction and as working environments demands higher scrutiny in detecting possible acoustical problems, which may create various difficulties, depending on the exact purpose of the room.

2.4.1 Room Modes

One immediate feature of small room acoustics is a result of prominent rooms modes - resonances of the room - which are sparsely spaced in frequency and in space at low frequencies. That may lead to a very different low and mid-frequency response in different positions in the room.

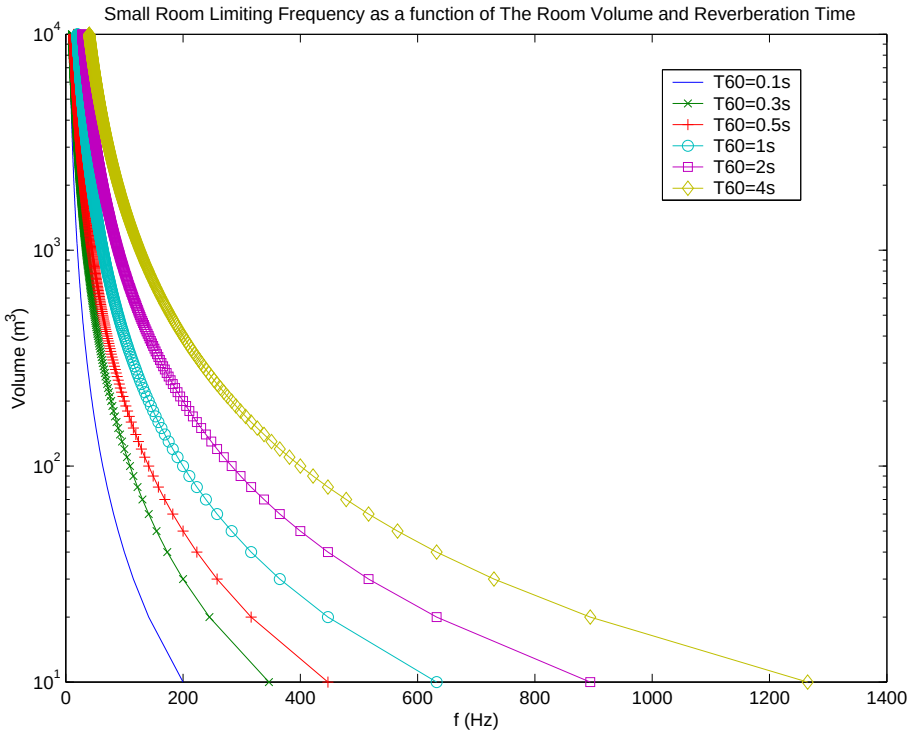


Figure 2.1: Room volume as a function of Schröder's frequency and the RT

As can be learned from eq. 2.9, two receivers placed in different locations in the room would necessarily hear differently the same source according to their specific dominant mode shapes. In large rooms and halls, these resonances are close enough in frequency and even though the frequency response at a particular point in the room may be very irregular, it is not normally perceived as a problem for our ears to smoothen out.

An instance of the frequency response variation at an arbitrary point in space in a room in shown in Figure 2.2. A difference as high as $40dB$ is possible in such curves, between the maxima and minima. Figure 2.3 shows a similar record of the sound pressure level, taken along a line in a room at a fixed frequency.

However, in small rooms, the case may be different if there are single resonance peaks (with little overlap from other modes), which are sharp enough (unattenuated, having a long decay time). Such resonances can be often heard and sound annoying when the a source frequency coincides with the mode frequency. In control rooms special equalization is sometimes done to suppress strong peaks or amplify dips resulting from such modes.

2.4.2 Early Reflections, Flutter Echo and Coloration

Room reflections arriving to a listener right after the direct sound are inevitable in any non-anechoic room, but can be perceived in different manners. It is usual to divide the reflection into early and late reflections, because of differences in their perception. Late reflections form the major portion of the reverberant field. It has a perceptual effect of enlivening and thickening the sound in the room, as opposed to a thin, localized and dead sound in an anechoic chamber, with zero RT.

In very large rooms a single reflection can sometimes be heard as an echo - a repetition of the source, separately heard by the listener as a distinct event, or a very late reflection. This effect can be very

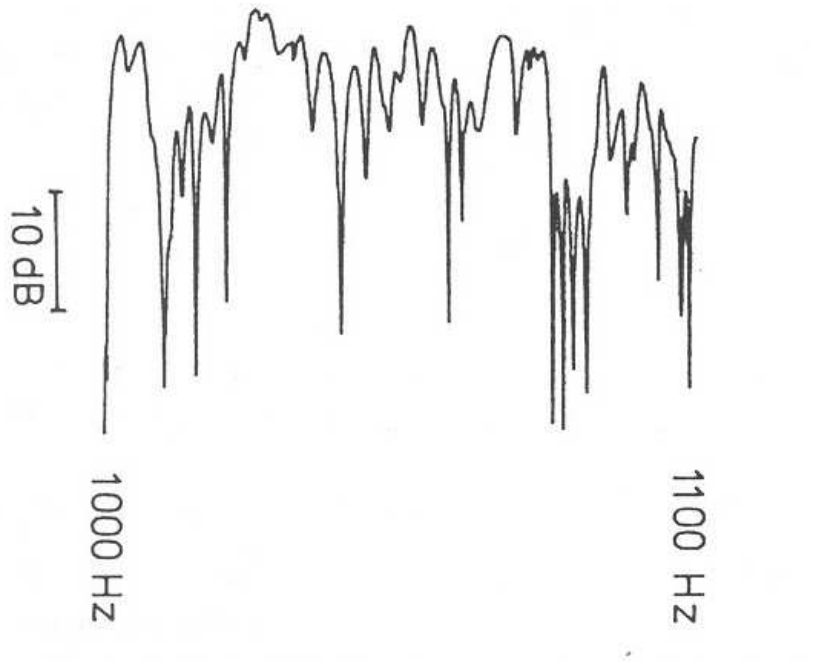


Figure 2.2: The logarithmic record of the sound pressure level of a point inside a lecture room at steady state (reprinted from [12]).

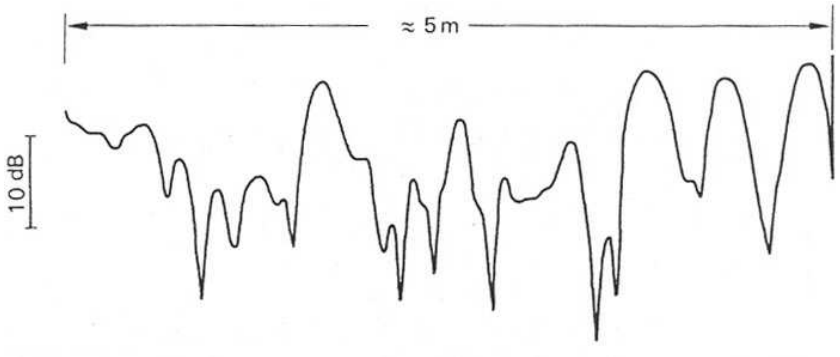


Figure 2.3: The logarithmic record of the sound pressure level along a line in a room at steady state (reprinted from [12]).

annoying, depending on the relative strength of the reflection and its time delay. Smaller rooms may still exhibit a flutter echo, where the repetitions heard are a series of reflections, still distinct yet not completely separated in time as the single echo. In case that the reflection appears very close in time to the direct sound, the auditory system can no longer separate the events temporally. That is referred to as the Haas Effect, named after H. Haas, who also found that critical time to be $20ms$. Below that threshold, separate events are perceived as one. In addition, there is an enhanced loudness perception of the original event. Still, this one perceived event is different spectrally than the original signal alone, as is shown below.

An impulse response, $g_1(t)$, of a direct sound and its reflection attenuated by a factor q and delayed by t_0 can be described as,

$$g_1(t) = \delta(t) + q\delta(t - t_0) \tag{2.17}$$

In the frequency domain, the squared absolute value of its Fourier Transform would be a comb filter of the form:

$$|G_1(f)|^2 = 1 + q^2 + 2qcos(2\pi f t_0) \tag{2.18}$$

An infinite series of reflections yields a similar filter, only having sharper peaks than with the single reflection.

$$g_2(t) = \sum_{n=0}^{\infty} q^n \delta(t - nt_0) \tag{2.19}$$

Then its squared absolute value is

$$|G_2(f)|^2 = [1 + q^2 - 2qcos(2\pi f t_0)]^{-1} \tag{2.20}$$

An illustration of the above in shown in Figure 2.4.

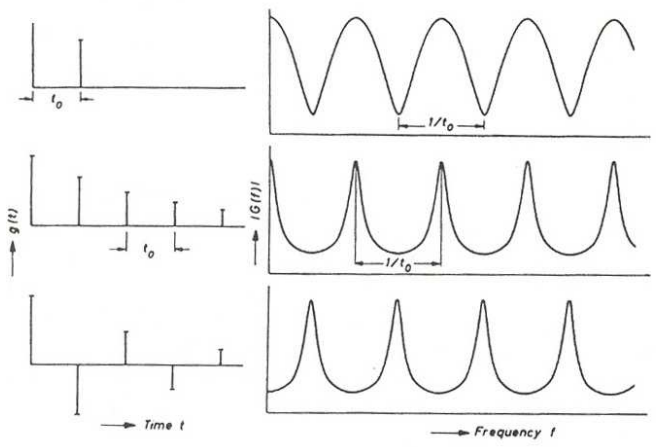


Figure 2.4: Impulse responses and absolute values of transfer functions of various comb filters (reprinted from [12]).

Coloration is generally referred to when the frequency response of the sound source is changed, when its output passes through an intermediate system, upon arrival to the receiver in a characteristic, periodic way. Namely, some frequencies would sound emphasized while other deemphasized, so that the source

image is distorted, but in a harmonic way. Such an effect may result from strong reflections, which create the above mentioned comb filter, or from strong standing waves between parallel walls, which also create a harmonic series of natural frequencies. These effects in small rooms, give rise to highly colored sound, and sometimes may be associated intuitively with a "boxy" sound. The implication of coloration may be noticeable depending on its severity. For instance, in control rooms the sound technicians make their decisions assuming that the reproduced sound, which is played by the control room's loudspeakers, sounds the same as the studio recording itself. In talk studios, over-emphasis of some frequency bands may give an oppressive or boxy impression to the listener and to the talker.

Sound coloration is difficult to estimate quantitatively, as it is first of all a qualitative impression that may or may not be observed by the listener. If it is noticed then there is a problem to assess the detectability and severity of a certain deviation from flat response. A completely flat response of the room transfer function is unrealistic (see figure 2.2 and 2.3) and there are unavoidable effects on the sound transmission.

A major source of coloration is room modes. In the mode-dominated frequency range where the modal density is low, it is imperative that the resultant transfer function of the room will become very non-flat.

There were a few attempts to find a quantitative measure for sound coloration. A quite complicated method is described in Kuttruff [12], using auto-correlation analysis of the impulse function of the room. It produces a measure of the randomness of the impulse response and hence indicates if there is any periodicity.

Rindel [16] suggested a Cepstrum analysis of the impulse response of a room, whereby the detection of periodicity in early reflections is facilitated, in some cases more accurately than with the former auto-correlation method. This method was not further researched.

A more recent technique for objective estimation of coloration was described by Meynial and Vuichard [17] in the context of the employment of active reverberation systems in auditoria. By assessing the difference between the measured sound field in a large room to the theoretical diffused field, which its frequency modulus follows Rayleigh distribution, they establish a measure for coloration. The validity of the distribution is true, though, only for large rooms (see eq. (2.16)) and could doubtfully be used correctly in small rooms.

It should be noted that a great deal of confusion in the coloration definitions was found in the surveyed literature. While some writers avoid using that term altogether [19], [20], [21] and [22], others use it freely and almost indiscriminately [9] and [18]. And yet others stick to the narrow definition presented above or restrict is in some manner, [16], [12], [15] and [17].

Olive and Toole performed a few benchmark experiments on the effects of both room modes [19] and early reflections [20] on timbre perception. They employed simulated sound field setups, with variable controls on the different reflections or resonances that accompany the direct sound. Both experiments suggest many important conclusions and thus only the most relevant ones are mentioned here. Some thresholds for resonance detection are established in respect to the program material - whether white or pink noise, music or speech - a single reflection with a certain center frequency and Q its delay time. They found in [19]:

1. white and pink noise are most suitable to resonance detection, more than music and speech. It is related to their dense broadband spectra and to being continuous signals.

2. Reverberation time has a detrimental effect on the resonance detection with discontinuous, impulsive and transient sounds. Modes with mid and low Q (broadband around their center frequency) are more easily audible in the presence of signal reflections and reverberation, compared to acoustically dead environment or headphones.

3. Single reflection detectability shows opposite trends, depending on the nature of the signal. With continuous broadband sounds at very short delay times have considerably reduced threshold than for longer delays (up to $25dB$ increase in threshold in the 0 to $60ms$ delay time range). Whereas,

4. transient sound delays are least audible at very short times, but the detection threshold decreases rapidly when they get longer (up to $40dB$ decrease in threshold in the 0 to $20ms$ delay time range). Generally high frequency resonances are easier to detect than low frequencies. Addition of reverberation to the signals changed this two-case distinction into one elevated threshold for all types of sounds.

In their sequel paper, which investigated the detection of early reflections, they report, amongst others, the following conclusions, many of which are similar in nature to the previous ones:

1. Delayed reflections from lateral directions are more audible than delayed sounds with direct incidence.

2. There are inherently different detection thresholds for reflection delays of continuous and impulsive sounds in a very damped or anechoic environment (or recording). With multiple reflections and reverberant field the impulsive sound threshold becomes the same as that for continuous sounds.

3. This reverberant field hardly affects the detection threshold of impulsive sounds reflections under $30ms$. With longer delays the threshold is gradually elevated. With continuous sound, the threshold resembles that of the elevated level under all listening conditions.

In another series of experiments Bech tested the impact of selected single reflections out of a group of early reflections on the timbral perception of perceived reproduced sounds, [21] and [22]. Using various constellations in a simulated standardized listening room, where the strength of single early reflections was carefully controlled, he divided the reflections into two groups: early reflections, with a delay time under $21 - 22ms$ - still under the Haas effect influence - and the later reflections with the reverberant field. His main conclusions were:

1. Only the first-order reflections, notably from the floor and ceiling contribute to perceptual timbral change of speech (discontinuous) signals. Noise (continuous signal) is affected by additional lateral reflections.

2. Increase in the reflection relative level is noticeable as well.

3. Filtered reflections, as of taking into account the room absorption, showed that the $500 - 2000Hz$ frequency range is detrimental for detection. Attenuated mid and high frequencies lead to elevated detection thresholds.

Other findings by Bech can be found in [24].

2.4.3 Image Shift

Another distorting effect of early single reflections is image mis-localization and changing of the perceived spatial impression and addition of "phantom images". This effect is especially disturbing for stereo reproduction, notably in control and listening rooms, where the stereo image is of particular importance. These effects are not dealt with here, but the reader may consult Bech [23], which is an additional part of the same series of experiments mentioned above.

2.4.4 Reverberation Time

Using the concept of reverberation time in small rooms at low frequencies is very problematic, because the definition for RT does not assume any mode domination, but an average of many decay times of all modes within a certain frequency band. At low modal density there is a very likely situation that the room is completely dominated by a strong resonance of one particular mode at a rather broad frequency band. The resultant RT would be composed from single mode decays, which defies the concept of RT

altogether, as was defined before. In some cases it would then be more accurate to talk only in terms of mode decay and not of RT, as the variations over frequency and position are very large.

The the use of RT is further complicated at very short decay times, when the sound field is very directional - may be composed from single reflections - and not diffused at all.

Another common feature of small rooms in particular, is having an unbalanced reverberation time curve, such as most curves used later in the presented experiments (figure 4.22, where the mid and high frequencies reverberate much less than the bass. That is, a longer decay time at low frequency bands than at mid and high bands. As the RT is an average of a number of room mode decay times, and modes are not easily suppressed at low frequencies, it is likely that they will resonate longer. The exact perceived effect of the unbalanced RT curve is unclear, especially when the RT is very short in small rooms. In large halls it is generally recommended to have a flat curve, although an elevated bass RT is recommended for classical music, which is expressed in a bass ratio value around 1.2, [25]. Naturally though, in most typical structures, there is limited absorption at low frequencies, which leads to the unbalanced situation. Therefore, it is rather difficult to design a room with flat RT over all frequency bands.

Recently it was shown that there is a reduced detection sensitivity of differences (difference limen) in short RT's compared to long RT's [26]. While for longer RT's the difference limen is around 4% of the total RT above $0.6s$, at short times of $0.2 - 0.6s$ it is up to 12% and below that it measured absolutely and is approximately $0.024s$. These values were measured for mean RT in the range of $250 - 4000Hz$, measured in octave bands.

2.5 Relevant Acoustical Measures of Small Enclosures for Subjective Characterization

In the previous section the main features of small rooms were reviewed: room modes, early reflections, coloration, image shift and reverberation time. Of all five, reverberation time is the most straightforwardly connected to a specific psychoacoustical perceptive measure, which is referred to as reverberance. The sensation of reverberance and its quantification by listeners is directly related to the reverberation time of an enclosure. This quantization becomes problematic at low frequencies, where the RT definition turns rather problematic, as was mentioned above (2.4.4).

A very useful concept, closely related to the RT in a room, is the Early Decay Time (EDT). It measures only the duration of the first $10dB$ of the decay curve of a steady sound source turned off, between 0 and $-10dB$, multiplied by 6. It was shown that the EDT is also highly correlated with the subjective reverberance perception in rooms (see section 3.4):

$$EDT = 6 \cdot t_{-10} \tag{2.21}$$

Here, as opposed to T_{10}, the measurement starts when the source is turned off.

The bass ratio is a quantity derived from the octave band RT values and can be related to the subjective brilliance and warmth of sound in concert halls.

$$BR = \frac{T_{60(125Hz)} + T_{60(250Hz)}}{T_{60(500Hz)} + T_{60(1000Hz)}} \tag{2.22}$$

The equation shows how balanced the RT between midrange and bass frequencies. BR of 1 indicates a flat curve. BR bigger than 1 indicates sustained bass compared to midrange and possibly treble bands, which can be perceived as lack of warmth and brilliance of the sound. BR smaller than 1 gives rise to an overall lack of perceived bass [25].

The case is different with the other features: their subjective quantification may be difficult absolutely, as well as relatively, and so any scaling by test subjects is either situation-specific, very localized in the room or completely unattainable. This is quite different than the case with large rooms, where various parameters derived from the statistical acoustical homogeneity of the room. Measures such as clarity

($C80$), lateral early reflections (LEF), strength ($G50$), definition ($D50$) and others are deemed useless for small rooms, which do not comply with the statistical assumptions for their definitions to be correct in large frequency ranges, have overall much shorter distances than in large rooms and thus could not be faithfully related to any specific listening sensation in the room.

Overall, it is clear that there is a difficulty to generalize the subjective performance of a room, or alternately to present it intuitively using compact figures and measures, not getting into complicated waterfall diagrams, reflectograms, impulse responses etc.

Chapter 3

Measurement Techniques

All experiments presented in the next chapters rely heavily on various acoustical and psychoacoustical measurement and experimental techniques at all stages. The techniques employed here are described in the following and include:

- reverberation time measurements through maximum-length-sequence (MLS)
- background noise measurements
- binaural recordings and reproduction and
- psychoacoustical tests

3.1 Reverberation Time Measurements Through Maximum-Length-Sequence (MLS)

Impulse function measurements are instrumental for acquisition of most room acoustics parameters. Through its impulse response it is possible to give a complete picture of the acoustics between a source and a receiver. Over the years, with the advance of technology, many methods were developed to obtain the impulse function of acoustical systems. These methods vary in their complexity, accuracy and susceptibility to noise.

The maximum-length-sequence (MLS) method became especially popular in the 1980's due to its faster algorithms with reduced computation time compared to other methods that were available. It uses a pseudo-random noise signal, with a $2^n - 1$ terms of binary predefined sequence of a certain duration. The missing term of "all 0" case is excluded from the sequence and yet it is very close to an FFT (Fast-Fourier Transform) complete block length. The autocorrelation function of this signal is very similar to Dirac's Delta function and thus, when the signal is repeated periodically as a pulse train, effective white noise is obtained, where all the frequency components are present with the same amplitude. In the processing of the recorded signal the MLS algorithm utilizes the Fast-Hadamard Transform (FHT), which is faster to compute than FFT. The MLS's drawbacks include vulnerability to distortion and time variance in the signal. In practice, a pulse train is not used and the MLS signal is passed through a *sinc* time window, with its associated losses, which have to be compensated. In addition, a pre-emphasis filter is normally harnessed to emphasize the bass frequencies of the MLS signal and in this way to compensate for some of the white noise shortcomings. A block diagram of the MLS with pre-emphasis is shown in Figure 3.1.

Nowadays, it is well-established that the use of deterministic excitation signals such as sine sweep is superior to MLS. Due to version limitations of the measurement system that was used in the present experiment (Acoustic Engineering's Dirac version 2.5), MLS was chosen as the best available measuring

method. For further details, the reader may consult Müller et al for a comprehensive review of MLS, [27].

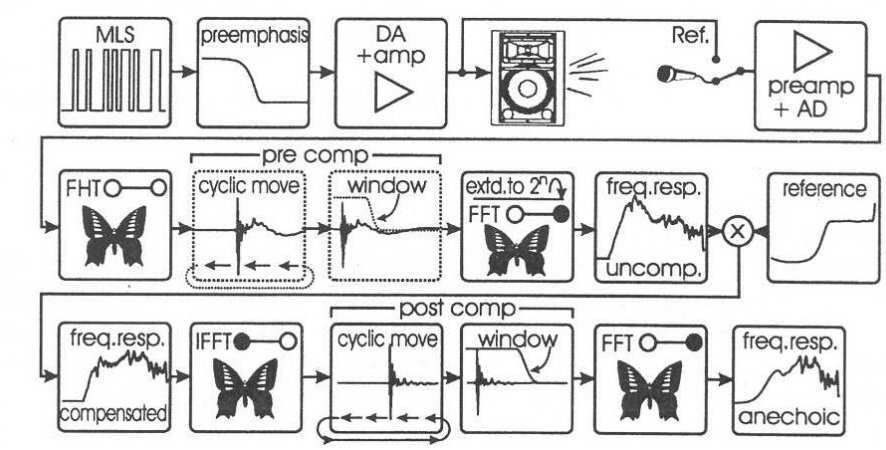


Figure 3.1: Maximum Length Sequence block diagram (reprinted from [27]).

Once the impulse function of a room is available, its decay curve, and hence its reverberation time, is rather simple to compute using Schröder's "method of integrated impulse response". This method gives the statistical ensemble average of all possible decay curves of the particular measuring conditions. Therefore it is much less susceptible to random errors which could contaminate the decay curve obtained [12].

In the present experiment three important points have to be taken care of: RT in small rooms, data acquisition at low frequency and the measurement uncertainties.

3.1.1 Small Room Specialties

As discussed in section (2.4.4), there is a problem of defining RT at low frequencies in small rooms, where some specific mode decays may dominate and puts a doubt on the interpretation of the RT as a solid quantity. On the other hand, there is no real alternative, but to average over all of these modes, as dealing with separate modes is both place-dependent within the room and is very cumbersome if not handled in octave or third-octave frequency bands.

3.1.2 Low Frequency Specialties

In addition to the above, recording of the low bass frequency data may be problematic. If either one of the speaker(s), amplifier, microphone and computer sound-card has an insufficient frequency response, which cuts off the low end of the bass, no valid data, corresponding to the noise level in the system will be recorded. Furthermore, in the frequency range lower than the lowest mode in the room, the energy is carried using that lowest mode, which has one decay time. As the farther apart from the lowest natural

frequency, the weaker is its level and measuring long enough a decay becomes very difficult, maintaining reasonable signal to noise ratio.

3.1.3 Measurement Uncertainties

The RT values that are obtained after at the end of the MLS process are not without error. The amount of possible error is dependent on the center frequency, the bandwidth and the number of positions measured. The relative standard deviation (in percents) is estimated using [28]:

$$\frac{\sigma(T_{20})}{T_{20}} = 88\sqrt{\frac{1+1.9/n}{NBT_{20}}}\% \tag{3.1}$$

$$\frac{\sigma(T_{30})}{T_{30}} = 55\sqrt{\frac{1+1.52/n}{NBT_{30}}}\% \tag{3.2}$$

where n is the number of decays measured in each position, N is the number of measurement positions and B is the bandwidth of the filter in Hz.

The nominal accuracy in the RT would be better than 5% for one-octave band data and 10% for one-third-octave bands, when $N = 6$ measurement positions are averaged, as was done in this experiment.

It can be seen that the uncertainty increases as the decay measured is smaller. And so it is expected that the uncertainty in EDT would be even higher than the RT, using only 10dB slope. However, the EDT computation needs only t_{-10}, as t_0 is obtained automatically by the definition of time zero (see eq. 2.15 and 2.21). This way the uncertainty in EDT is smaller by a factor of $\sqrt{2}$ compared to T_{10} assuming equal error for each t_x. Thus it may be given by:

$$\frac{\sigma(EDT)}{EDT} = 175\sqrt{\frac{1+3.32/n}{2NBT_{20}}}\% \tag{3.3}$$

3.2 Background Noise Measurements

Ambient noise sources in small rooms can be either internal - such as cooling fans, heaters and air conditioning systems - or external, such as passing traffic and people's activity in the building. Depending on the exact use of the room, in general, the background noise in the room need be neither annoying nor intrusive. If music recordings or performances are to be made in the room, the noise must not mask the lowest sounding musical details. Tonal imbalance, periodicity, unsteadiness and vibrational components all markedly increase the potential annoyance from the noise.

As the human ear is not equally sensitive to frequencies in different bands, it is normal to weight the sound level measurements in some manner. A-weighting is adequate for some purposes, but gives only one number in dB(A), which can describe many different noise spectra that may in turn exhibit very different annoyances.

A standardized and popular for measuring indoor noise is the NC noise rating, first described in [29]. It sets a stricter criterion to the noise rating, by noting any deviation from a set of reference curves, even at one frequency band. This method suffers from a similar problem as the simple A-weighting, which is its inability to discern differently sounding spectra and rating them with the same NR.

Somewhat similar method is also used that overcomes this problem. The RC noise rating gives the rating along with a letter which designated the tonal (im)balance quality of the noise [30] by Neutral (N), Hissy (H), Rumbly (R) or Perceptible Vibration (RV).

The NR rating method is the one normally referred to in listening room standards such as [6] and [7] and in other recommendations for background noise levels in rooms. Therefore, NR method was chosen as the primary method for rating the background noise in the rooms in the experiment, yet RC ratings were calculated and presented as well.

3.3 Binaural Recording

The human sensory auditory information can be reduced acoustically to two signals arriving at the ear drums in each ear. Reproduction which is able to replicate these signals should hence give a complete image of a sound field, both timbrally and spatially. Binaural recording technique utilizes two microphones at the entrance of the ear canals of a human head or a dummy head and records the signals into two stereo channels. Playing back the recordings in headphones produces a virtual impression of the sound field, as would be experienced had the listener been at the original location - spatial perception included.

In reality, this straightforward procedure proves to be much more complex and its successful replication of the real sound field is currently still limited.

The central principle behind the binaural technique is to provide all the spatial cues for auditory localization in the recording. These include time delay and level differences between the two ears, and spectral cues mostly produced by the special shape of the pinna (the outer ear), which behaves as a sophisticated filter. By analyzing the two signal streams from each ear, the brain is able to localize the source position in relation to the listener with remarkable accuracy. In addition, head movements create small changes in the received sound in both ears, which can further aid the localization if there is some uncertainty the source location. Finally, the brain uses the information from other senses, notably vision, to complete and complement the picture of spatial events.

The ratio between the free-field sound, measured a certain position without the head and the sound field at the same point but in the ear canal, after the outer ear filtration is added and including the reflections from the head and body, is called the Head-Related Transfer Function (HRTF). This function is individual for every person. At frequencies below $1.5 - 2.5kHz$ it is remarkably similar for most people. Above, there are increasing differences between individuals, as is reported by Møller et al, [31]. See Figure 3.2 for an example of the variations between subjects.

Repeating, the success of binaural recording would be to provide a potential listener with exactly the same signals one would hear had he actually been at the recording time and place, replicating the same sensation as if being present there for real. The difficulty arises as there are no two identical pinnae. Every individual is trained to interpret the sound field using one's own pinnae or HRTF. Differences in the pinna shape and ear canal length mean a different spectral image arriving to the eardrums. Hence, outer ears are not interchangeable, as every person hears a unique image, resulting from unique filtering.

Similarly, using circumaural headphones, which cover the outer ear completely, is shown to have individual frequency response. Thus, the equalization of the headphones is not straightforward and ideally should be done individually, by measuring the Headphone Transfer Function (referred to as HpTF or PTF), [32], [33] and [34].

The implication on binaural recordings is crucial if spatial localization is to be maintained. In order for an individual to experience precise emulation of the sound field, the recording has to be made through one's own ears, placing the microphones at the ear canals' entrances. Then, the responses of the microphones and later of the headphones have to be equalized to produce flat response against any tonal imbalance due to the transducers. Furthermore, the head movements of the listener wearing the headphones should preferably be tracked, so that the played-back signals can be corrected in real-time, giving the impression of a true presence in the sound field. This so-called individualized recording is in most cases impractical. A major challenge in designing artificial pinnae for dummy heads is to find a pair of "golden ears", for which localization cues would still be perceived correctly for most listeners. In a series of experiments it was shown that localization ability from binaural recordings, done with a variety of dummy heads, still falls far from the ability to localize correctly in reality.

However, there are a few steps which can be performed to aid the localization and the reproduction of the recordings. First, an attempt should be made to record through ears that show improved results compared to other ears. It appears that some ears are more universal and introduce an HRTF, which can be interpreted correctly by a larger percentage of the population. Ear molds used in the dummy heads are such an attempt. In addition, equalizations of the headphones and the microphones used have to be

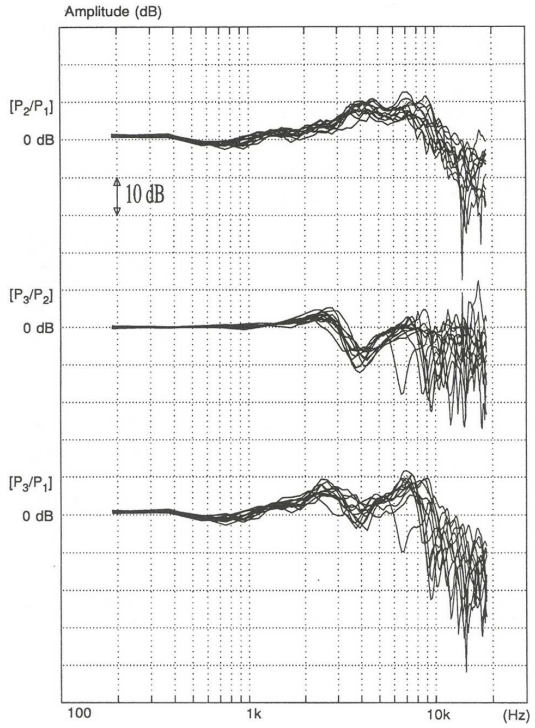


Figure 3.2: Various measured transfer functions of 12 human subjects' left ears for source positioned at the left side of the head. **Top:** HRTF; **Middle:** ear canal pressure divisions (independent of direction); **Bottom:** the resultant combined transfer function; (Reprinted from [31]).

done. The headphones equalization can be performed either individually or non-individually, quite like with the recording itself. If the non-individualized approach is taken, once again, a generality of the HpTF should be achieved, so the particular equalization of the headphones will be valid for as many listeners as possible.

Special attention should be drawn to a repeatedly reported mis-localization type, that which is on the median plane. The median plane is the on-axis perpendicular plane to the head, which connects front, above and back (Figure 3.3. Even in reality most confusion is found on this plane, when stimuli from the back are mistaken for front or above and vice versa (Figure 3.4).

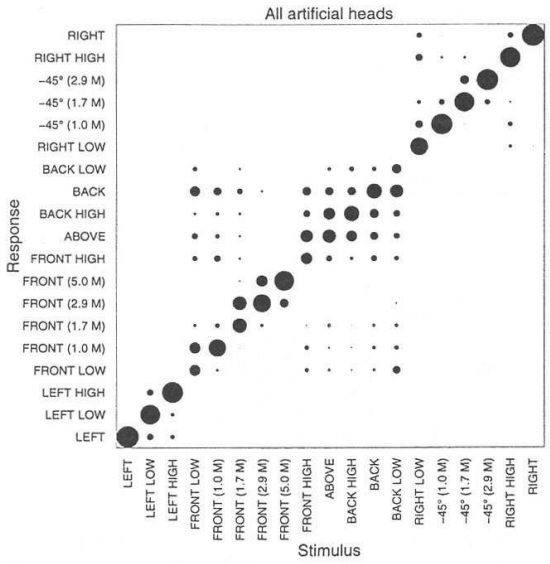


Figure 3.3: A collection of mis-localization mistakes for 9 different dummy heads. The number of answers related to a specific pair of stimulus/response is marked by the point size (8 subjects, with 912 stimuli per subject). (Reprinted from [35]).

More references on binaural recordings can be found in [36], [37] and [38].

3.4 Psychoacoustical Tests

The final goal of many room acoustics measurements is to find the way sound in a given room is perceived by people who hear it. Perception of sound is a very general term and could mean anything from detectability to preference of certain sounds, their thresholds, variations etc. Psychoacoustics works with the physical stimulus and the consequential subject's response, which may be affected both by physiological and psychological processes of the subject [39]. In such an experiment as is presented here, the tools provided by psychoacoustics are employed in order to correlate physical phenomena of the sound field, to a certain sensation experienced by the listener.

Once the possibilities apparent from the available tools and material of the recordings are determined, or at least assessed, then the relevant psychoacoustical methods have to be chosen in order to realize that possibility with the highest confidence level, maintaining practical terms.

In this section the psychoacoustical testing methods used in this project are explained.

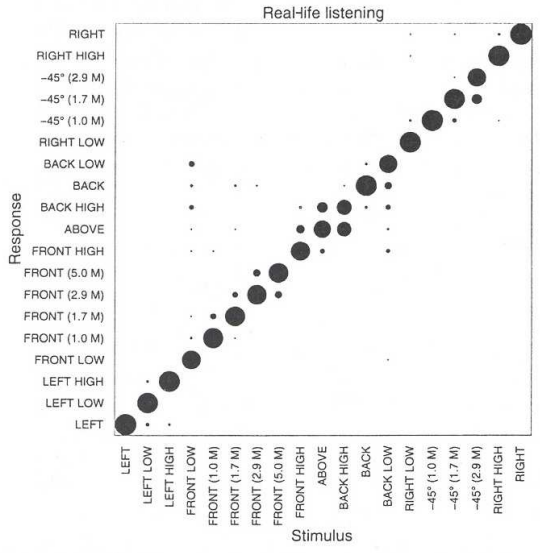


Figure 3.4: A collection of mis-localization mistakes in real-life listening. The number of answers related to a specific pair of stimulus/response is marked by the point size(20 subjects, with 114 stimuli per subject). Reprinted from [35].

3.4.1 The Rating Scale Method

The rating scale method have enjoyed much popularity in a wide array of psychological tests, mostly due to the ease of implementation and their appeal in evaluation of human traits, human reactions and their reactions to various stimuli. The variations of this method are many and only the one employed in this experiment is explained. Acoustical implementations are sometimes based on Poulsen [40], but most of the following discussion is based on Guilford [41].[1]

In the general procedure of a rating-scale test a subject is asked to give a rating to a certain thing. The subject is given a scale, which is designated by literal descriptors, which mark clear points of the numerical or graphical scale. The test subject is to be given maximum indication by the literal ratings of the exact nature of each step in the scale, hopefully with minimum ambiguity and bias, due to ill-chosen descriptors. The premise of the test relies heavily on the ability of individual to give a rating to a certain aspect of the test object.

The problems that may be involved in implementing this relative simple method are numerous and special care must be taken to avoid in advance as many difficulties as possible, by correctly constructing the test.

Scale Size

The choice of the particular scale size employed in each test is dependent on the ability of subjects to discriminate between the different scale steps, which in turn depends on the particular rated object. A

[1] Guilford's classic book "Psychometric Methods" from 1954 was used extenisively here as a major reference throughout both construction and analysis of the tests. Most information in it is not dated, despite its age and most ideas and statistical techniques can be applied more easily today using modern computer technology. More modern and less conventional methods of course exist, but were not surveyed here.

clear and daily trait might be considerably easier to grade than a more esoteric feature, whose exact level is more difficult to pin down. In practice it means that the easier the object is to relate to, the more scale steps can be used. However, there is a limit, and probably an optimum, to the level of discrimination, even with very clear knowledge of the scale areas.

Upon constructing the scale, several common rater behaviors should be considered. First, raters tend to avoid extremes. For instance, a 7-scale may be effectively be used only as a 5-scale by some raters ("central tendency error", see below). Second, too small a scale may be too coarse an estimation for a given trait. A fine scale will depend much on the willingness, motivation and capability of subjects to distinguish between the different scale steps. Otherwise, there would be reliability problems in the ratings. Third, the ability to use a finer scale depends on the subject training.

A scale can be either unipolar, or bipolar where the object can be rated both in terms of deficiency and excess of a certain trait. There is no rule of the right number of scale steps to be applied. Nonetheless, the prevalent 5-scale and 7-scale options for the unipolar test seem to be a rather safe and solid choice.

Scale Descriptors

The literal descriptors of a trait should be unequivocal and lingually familiar. They should be chosen so that they describe gradually and rather continuously a changing feature, which fits well with the numerical or graphical scale. Finally, the words used must not imply a prescribed opinion of the trait. For instance, a tag saying that a recording is "awfully boomy", in the test presented here, would suggest that a boomy recording is a bad thing, perhaps opposed to what some people feel.

Common Errors

The rating tests assume a human capability to give quantitative observations. However, people are subjects to many biases and often cannot give objective observations, rendering the test an imperfect measurement method. There are quite a few errors possible, which may affect the results acquired through a rating-scale test.

- Leniency Error - subjects tend to rate differently familiar objects. For example, when musical excerpts are played back, familiar program may incur either higher or lower ratings, according to the subject's preference. An obvious way to tackle it, would be to use unfamiliar programs if possible.

- Central Tendency - as mentioned above, subjects tend to refrain from using the scale extremes. One reason can be due to the harsh wording of the edges, which might seem indeed too extreme. Another may be a lack of knowledge of the range of variations to be presented. Thus, a subject will have to adapt oneself to the degrees encountered throughout the test. As the a test is revealed little at a time, the subject relative "scale scaling" is performed as the test goes along (the so-called learning in scale formation).

- Logical Error - happens when the phrasing is ambiguous (see above).

- The Halo Effect - Subject tendency to overrate or underrate traits according to their overall impression of the stimulus. In this case, overvaluing or undervaluing boominess, boxiness or overall sound quality ratings differently for different programs. An extra inclination for that error can happen due to the rarely discussed and defined and not easily observable nature of these traits. A side-effect of this error is increase in the apparent intecorrelation between ratings. Separation of questions (having a single question after each stimulus) is an effective measure to tackle this intercorrleation. It was not employed in this case, due to a severe extension of the test time to be needed. Instead, an extra attention was paid to the scale construction. The three scales are unequal in size, use completely different phrasings and do not suggest of one another, as will be shown later in section 6.3.1.

- Proximity Error - rating of features which is done close in space and time tend to intercorrelate more highly than had it been done separately. Like in the halo effect a separation of the ratings would help to avoid this error and intercorrelation between different ratings seems even less likely. Even so, it might be important to treat any intercorrelation found between two traits with caution.

Many errors can be minimized by training of test subjects by familiarizing them with the tasks involved and the type of stimuli to be handled with in the tests. Repetition of tests occasionally show a learning effect, as subjects improve or show better repeatability of identical tasks.

3.4.2 Double-Blind Quadruple-Stimulus with Hidden Matching Pair

The other method that was harnessed in that work had been chosen more organically and does not comply intimately with an existing method. The closest popular similar test would be the double-blind triple-stimulus with hidden reference method, used extensively in small impairment detection tests, which are described in the ITU standard [6]. In analogy, the system employed here can be named double-blind quadruple-stimulus method. The subject is presented with the task of matching two samples that were recorded at the same room, but at different positions within the rooms. The sample A is to be matched with another hidden sample B,C or D, whose exact location is randomized at every run of the test. It is implemented as a computerized test, where the playback of each sample, the randomization process, both of the questions and the hidden matching pair and the validation of correct answers are completely automated. The test subject is forced to choose between the alternate possible answers.

A more complete description of the system's details is given later in section 5.4.

3.4.3 Ecological Acoustics Methodology

A variation on the previous test was the final test performed by the subjects. The task and interface are identical to the triple-blind quadruple-stimulus described above, but this time the four samples were four different programs - both music and speech. The task demands much more from the subjects and puts to test their 'capability to discriminate between rooms independently of the program played.

The underlying concept behind this task is suggestive of the experimental principles normally used in the field of ecological psychophysics and notably ecological acoustics. Specifically to the general trend to vary more than one parameter at a time in experiments which put to test auditory abilities that are normally used in daily life - so-called ecological tasks. For example, a typical experiment would be asking a test subject to judge what is a bar's length, according to its impact sound after being dropped on the floor. The bars presented are varied in length, but also in material. This principle defies traditional experimental design, in which parameters are separated and varied independently where possible. Moreover, it makes use of sounds which are more complex by nature than various psychoacoustical test sounds such as pure tones. Ecological acoustics therefore assumes more far-reaching capabilities of the human analysis of sounds, and tends to treat them more holistically than conservative psychoacoustics. An introduction and further references to this field can be found in Gaver [42].

The variation described above above does not fall into the ecological acoustics slot per se, mostly due to the inorganic nature of the samples (ecological acoustics excludes speech and music from everyday sounds). And yet it hypothesizes that the room acoustics is audible between very different conditions: different positions within a room and different test samples. In practice, the subject has to "hear through" modest changes in playback levels and left-right channel balance as well and detect the same room acoustics.

3.4.4 Experimental Design

There are certain factors in human psychophysical tests, which have to be taken into account, lest they would adversely affect the test results. Two are discussed here in conjunction with the above mentioned tests.

Sequential Bias and Latin Squares

Unwanted influences on test results, or bias, can be the consequence of various elements. A sequential bias may occur when a test subject performs an evaluation in a way that is influenced by the previous evaluation he did. A common way to reduce this bias would be to change the presented sequence of stimuli to be evaluated between test subjects and repetitions. A typical design would thus employ a systematic plan for the stimuli order determination such as a balanced Latin square [40]. In a Latin square design every row of the square represents a sequence of stimuli or questions. In Latin squares every number appears only once in every column and row of the square. Using balanced Latin squares also ensures that every two stimuli are consecutive only once in the experiment (twice in case of odd-numbered squares). The square is constructed according to the characteristic row:

$$1, 2, n, 3, n-1, 4, n-2... \tag{3.4}$$

Where n is the number of rows. In order to construct the following rows, 1 should be added to each term of the previous row (modulus n).

Fatigue

Generally speaking the longer a test is, the more reliable are the obtained results. Nonetheless, too long a test can be very demanding and thus cause untimely subject fatigue. Therefore, it is necessary to find a suitable length for the test, which incorporates appropriate breaks in between test parts, when the subject may rest.

3.4.5 Statistical Analysis

The eventual analysis of the measurements involves the use of many statistical concepts and techniques where needed. The use is not always entirely rigorous, but an effort was made to bring it at least to a point where some confidence (or qualification) can be obtained regarding any subsequent conclusions drawn from the data. The following references were used for the statistical background: Guilford [41], Bartz [43], Wuensch [44], Dorak [45] and [46]. Matlab software was used for all statistical analysis and its technical documentation was regularly used as well [47].

3.5 Hearing Threshold Measurements

As a prerequisite for the tests, all subjects must have normal hearing. Those who were not recently tested for normal hearing were tested before approaching the experiments. The method used was the ascending method, performed manually. Audiometer AA222 by Intracoustics was calibrated to Sennheiser HDA200 audiometric headphones. The ISO 8253-1 pure tone audiometry standard was followed for the measurement. Consult Poulsen [40] for further details.

Chapter 4

Experimental Stages

Two main stages of the experiment, which are covered in this chapter:

- recording equipment and program selection,
- room data acquisition: reverberation time measurements and program recording

The subsequent listening test construction is covered in the next chapter.

4.1 Recording and Playback Equipment Selection

Various setups were tested to obtain the best quality out of the samples recorded in the different rooms.

Each one of the intermediate stages can lead to inferior quality if badly chosen. It is the intent of this experiment to have the recording and playback chains optimized to produce as high fidelity sound output as possible. This way secures that any later preference or rating of the recorded material by the test subjects is not attributed to the equipment and recording methods, but only to the contents and to the exact questions asked. Ideally, the only two variables in the above chains would be the room and the listener.

4.1.1 Recording Chain

The recording chain that was eventually chosen and used is illustrated in block diagram in figure 4.1 and is described in detail in the following.

The Source Programs

Initially the program material was chosen to be white noise, anechoicly recorded speech and possibly music, all to be reproduced monophonically. The different programs can be utilized to examine the different attributes of the small room acoustics, apart from image shift, which is not dealt with in this experiment. White and pink noise are useful for coloration detection, while speech may be adequate for reverberation time changes (see section 2.4.2 for more details). Anechoic recordings of speech are available in good quality. White noise can be easily produced by a noise generator. Any music material used should not be anechoic, if a realistic situation is to be simulated, where people listen to recorded music, which has its own specific reverberation. Studio recorded tracks are not anechoic, but "breathe", according to the type of music and its production. Anechoic music recordings can be used to simulate the situation where real musical instruments are played in a room. But most instruments have highly complicated radiation patterns, which cannot be emulated by a normal loudspeaker. Therefore the this option was not pursued here.

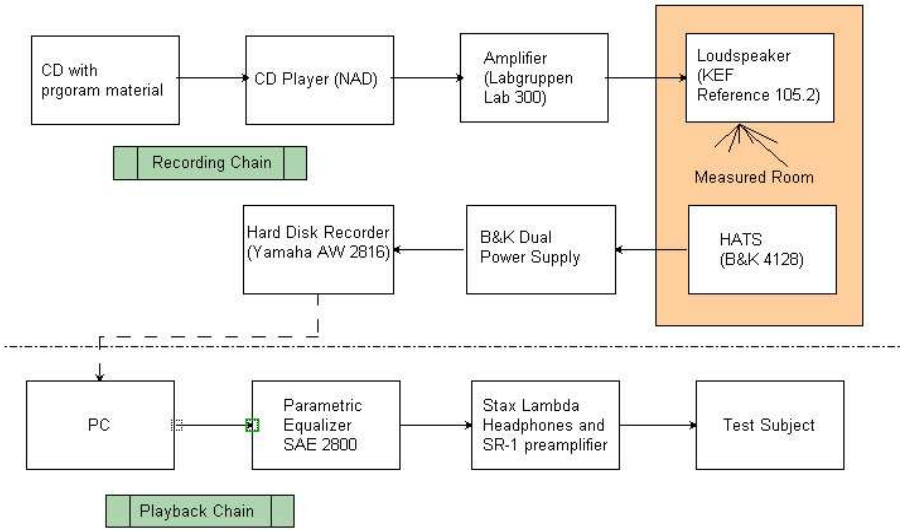


Figure 4.1: Recording and playback chains block diagram

All the samples used were recorded on one CD and normalized to have more or less the same loudness. It was done using Adobe's Audition software. This normalization may pose a problem of degradation in the dynamic range of some recordings, while elevating the noise floor of others (but keeping the signal to noise ratio). As a rule, the normalization must not show any clipping, which is then often heard as digital distortion.

The music samples were chosen first according to their original recording fidelity, which should be as high as can be found. Then the excerpts should still show good quality in either one of the monophonical reproduction possibilities: right, left or mono. The latter was never used. In addition, although still far from complete, a wide variety of styles is represented: medieval, chamber/world, electronic, funk, jazz and rock music. Finally, the instrumentation in each sample was unique and involved some changes within the excerpt running time, which normally involves instruments coming in and out.

The most important question is how well can a certain sample be used in order to pinpoint the various room acoustics features. As the compilation included both very produced material and live, acoustic recordings it is likely that at least a part of them would end up being useful for the sound quality evaluation. The inclusion of electronic music is, in a way, a shot in the dark. It contains no familiar instruments (apart from female vocalist), but only sampled or synthesized sounds, which can be referenced only to themselves and not to any idea that may priorly exist in the listener's mind.

The following samples were initially compiled on the CD. All samples are $15 - 21s$ long:

1. **White noise**

2. **Pink noise**

3. **Female anechoic speech** (B&O CD, which was recorded in the Acoustics Laboratory, DTU, 1992 for the Archimedes project))

4. **Male anechoic speech** (B&O CD)

5. **"7 Anis" - Egberto Gismonti** (taken from "Infancia", 1990, ECM). Grand piano, cello and double bass.

6. **"Puis qu'en oubli sui de vous"**, rondeau by Guillaume De Machaut, performed by the Ferrara Ensemble (taken from "Mercy Ou Mort - Chansons & motets d'amour", Arcana/WDR, 2001). Tenor and bass singers and vielle.

7. **"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.

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

9. **"Mode E - Single Solos and Group Dance"**, Charlie Mingus (taken from "The Black Saint and The Sinner Lady", Impulse, 1963). Piano, bass, classic guitar, brass section and flute.

10. **"Soul Catcher", The The** (taken from "NakedSelf", Nothing, 2000). Acoustic guitars, electric guitars, bass and male vocals.

Two more aspects were considered before and after the tracks were chosen. First, the music should not arise any subjective like or dislike. Second, the tracks should not be familiar to the subjects. Both considerations are necessary to reduce unwanted biases, such as those discussed in section 3.4.1.

The 16-bit comiplation CD with all 10 tracks was played using a NAD compact disk player.

The Amplifier

A commercially available amplifier which is reasonably linear can be found easily. The Labgruppen Lab 300 amplifier, which was also used for the room acoustics measurements, was used for the playback. It is both powerful, portable and has excellent fidelity [48].

The Loudspeaker

Selecting an appropriate loudspeaker can pose some problems. The first consideration is a broad frequency range, which covers easily the low frequency range, which is of particular interest in the experiment. Second, flat frequency response and high fidelity are also important for correct reproduction. Third, frequently big speakers which cover the desired frequency range may not sound natural for signals such as speech. They may also excite extra modes in the room, due to their size. In the case of speech reproduction, the recording might not sound natural, as the source size and its radiation pattern differ from a real person. Nonetheless, one aim of this experiment is investigating these excitations of different room modes and the resulting coloration of signals in various room acoustics.

A few available loudspeakers were subjectively tested, in order to choose which one should be used: two units of KEF 105.2 Reference and Quad electrostatic speaker. Despite having a very smooth sound, the electrostatic speaker was not an option, due to its very broad radiated image and the uncommon radiation pattern (figure of eight), which is not encountered normally in small rooms. The two KEF speakers were not identical and the preferable one was chosen in an informal subjective test, as it showed less tonal imbalance.

The loudspeaker has one separate box for the woofer and another for the midrange and the tweeter. The bass response extends down to $38Hz$ flat, according to the manufacturer. Since the speaker is not new its response was measured in the small anechoic chamber in building 354, Ørested. The relative response was recorded using B&K Pulse system.

At any rate, since the low frequency response is instrumental for the experiment conclusiveness regarding low frequencies, this model was favored over some other smaller yet good speakers.

The transfer function of the speaker can hardly be considered flat, but it shows a $\pm 3dB$ between $50 - 3000Hz$. Nonetheless, its fidelity sounds satisfactory and there was no real alternative, with comparable bass response. At around $3.5kHz$ a dip in the response can be seen. It is likely that it was intentionally

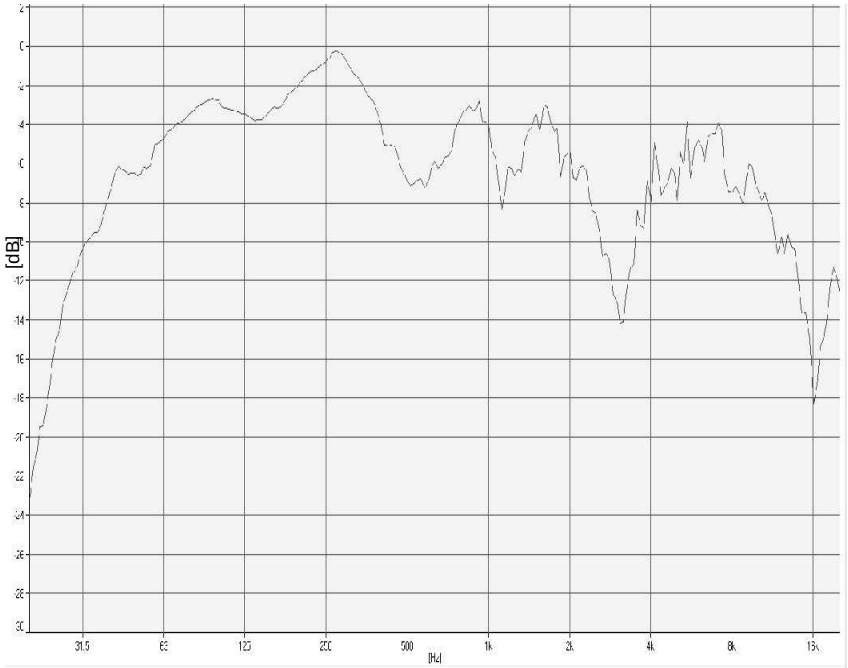


Figure 4.2: The on-axis frequency response of the KEF Reference 105.2, as was measured in the small anechoic chamber, building 354. Gridlines show 2dB intervals on the ordinate.

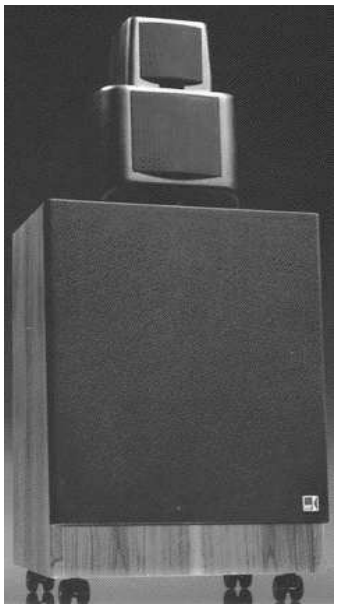


Figure 4.3: The KEF Reference 105.2 loudspeaker

designed so, in order to compensate for the ear canal amplification resonance at this frequency. Off-axis the high frequencies above $2000Hz$ drop slowly relative to the on-axis response, but the function's general shape is maintained.

Recording Microphone

As the recording is to simulate small room listening, it was decided to record binaurally, using a dummy head. With the initial lack of a proper head and torso simulator, a substitute was tested - a Sennheiser head simulator type MZK, with compatible microphone headset type MKE 2002. It can be mounted on a tripod and have the matching microphone pair worn on the head as a headset. The microphones are positioned at the entrance to the ear canal, just outside of the outer ear. They are directed upwards and connected to a DC voltage supply. There are three immediately noticeable shortcomings with this solution. First, the position of the microphones is close to the ear canals, but not quite in the realistic location and thus inaccurate. Second, the dummy head is made from hard, lightweight material, which does not simulate accurately real head. Also the shoulders, which add something to the diffracted and reflected sound at low frequencies, are missing. And last, the microphone frequency response are not provided by the manufacturer and their fidelity is thus questionable. Of course the latter can be measured, in case that preliminary tests prove it to be of high quality after all. Nonetheless, the quality of the initial test was not satisfactory and it was decided to obtain a Head and Torso Simulator after all.

The B&K Head and Torso Simulator (HATS) type 4128 was employed for all binaural recordings. It contains ear simulator with a $1/4''$ microphone behind the artificial pinna. There are no ear canals in the ears, so no correction has to be made to correct for their effect later (as subjects will get the signals as if heard from the entrance to their own ear canals). Its picture is shown in figure 4.4. The HATS is built a half human body with a torso, wearing a vest, which is not shown in the picture.

Initial tests in the anechoic chamber showed a very good sound fidelity with mixed localization performance, depending on the incidence direction. It complied with reported performance that was accounted in section 3.3. Similar performance figure was reported specifically for the 4128 HATS and is shown in figure 4.5.

Figure 4.4: The B&K head and torso simulator type 4128

The Recorder

The last stage in the recording chain, the recorder, is of crucial importance. In fact the weakest point is the input amplifier for the microphones (such as a mixer). While most recorders are able to produce very good results for line level recordings, the amplification for low level inputs such as microphones, is in many cases of noticeably poorer quality. It can result in murky, colored and noisy recordings, even when high quality, professional microphones are used.

Digital recording ensures compatibility between the recorder and the playback and no further degradation - at least from the first two stages of the playback chain - once the sample has been recorded. The

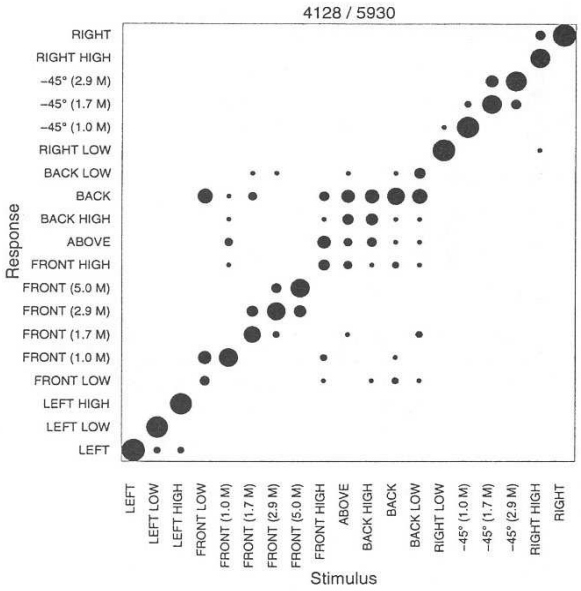


Figure 4.5: A collection of mis-localization mistakes for the B&K 4128 head and torso simulator. The number of answers related to a specific pair of stimulus/response is marked by the point size (20 subjects, with total of 380 stimuli). (Reprinted from [35]

case is different if the recording is analogue and has to be transfered (converted to digital, for instance, to be computerized later) several times before reaching the listener's ears.

The hard disk recorder (HDR) by Yamaha, type AW 2816 (figure 4.6), was employed due to the convenience of use and the option to later export recorded samples to WAV format or to burn them directly to an audio CD. It includes two analogue microphone inputs, enough for the HATS, which are then converted to digital inside the machine. All controls apart from the gain trimmer at the inputs are digitized. The HDR can operate in either 16 or 24 bits. Unfortunately the 24 bits could not be exported easily out of the machine, and all recordings were done on 16 bits. [1] Even so the recordings enjoy a high reported dynamic range around $100dB$, with low noise floor, sounding remarkably quiet.

4.1.2 Playback Chain

The playback chain is shown above in figure 4.1. All of the above regarding the recording equipment high fidelity is relevant for the playback chain as well. The most complicated part is the binaural reproduction through headphones if it is to give an impression of a true virtual room.

The Recorded Media, Playback Machine and Amplifier

The Recorded samples were recorded digitally and thus do not degrade when manipulated and repeatedly played back. The playback quality should match the recorder playback capability, using a decent digital to analogue 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 sound card used for playback.

[1] Only later was it discovered that the option was actually available all the time, yet was not noticed.

Figure 4.6: The Yamaha AW 2816 hard disk recorder

The Headphones

In attempt to choose headphones that are as accurate and close to the source as available, the Stax Lambda were chosen over Sennheiser HD 500, HD 600 and Sennheiser HD 250 Linear, through an informal subjective listening test, which is described below. The Stax Lambda circumaural electrostatic headphones come with a special class A electron tube preamplifier (Stax SR-1), which also provides the polarization voltage needed for the headphones to operate. This model is considered a very high-quality one and has an excellent sound. Figures 4.7 show previously reported PTF's and their respective mean and standard deviation. In the quoted research [33]) 14 types of headphones were compared and their PTF's were measured. The Stax Lambda showed a relatively small spread and variations between test subjects. Even so, it is clear that the PTF's are by no means identical for all the test subjects and that as the frequency increases so do the differences between the subjects.

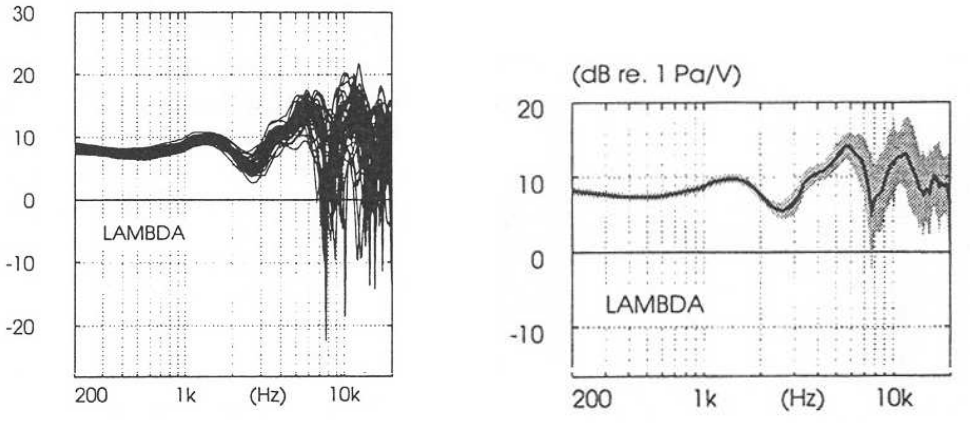


Figure 4.7: Left - Stax Lambda electrostatic headphones reported Headphone Transfer Functions (PTF's) measured on 40 subjects; right - Stax Lambda electrostatic headphones reported mean frequency response, with its standard deviation (reprinted from [33]).

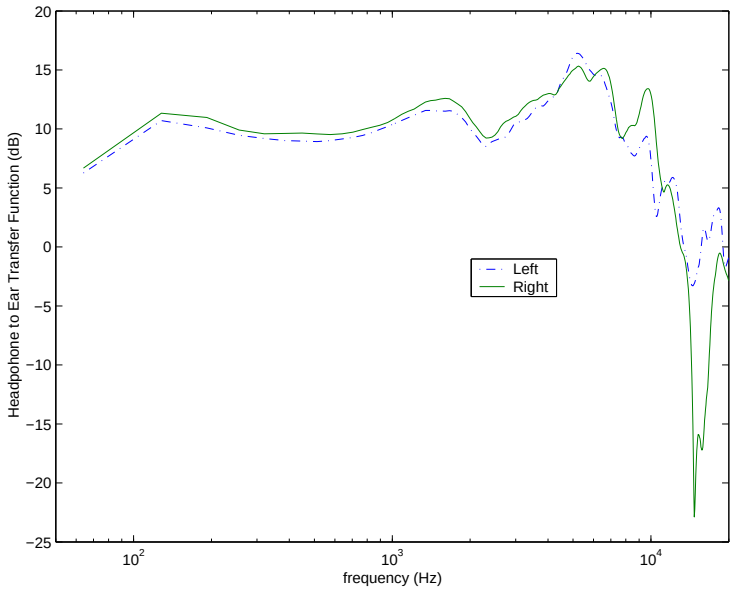


Figure 4.8: Stax Lambda electrostatic headphones measured Headphone Transfer Functions (PTFs) measured on B&K HATS 4128

A 20s music sample was recorded with the Yamaha HDR and with the Tascam portable DAT recorder and the B&K dummy head, both in a small room and in the large anechoic chamber. The recording was played-back, using all headphone types. The Lambda sounded the closest to the original, from the various headphones. There was a great similarity between the two recorders, but the Yamaha HDR seemed somewhat superior.

There were two immediate problems detected in these test recordings. The first was overemphasis of the small room acoustics. The second was in-head sense of median plane (front-back), for recording in the anechoic chamber. The two problems are likely to stem from the same source, namely, non-individual head transfer function used with the HATS and non-equalized headphone transfer function. Only the latter can be addressed with the available means.

The Equalizer

Even with high quality headphones there are still significant differences in the way different subjects spatially perceive the binaural recordings. It is generally established that equalization of the headphones to microphone transfer function, obtaining a flat response, is the minimum which has to be done, in order to obtain the so-called non-individualized reproductions. Individualized reproduction will take place if such equalization is done individually, using the subject's ears.

In order to obtain a binaural recording free of the frequency response of reproducing headphones their response has to be measured on real ears using a miniature microphone at the entrance to the ear canal of each test subject (individualized transfer function), as described by Møller et. al. With the lack of the proper microphones and the tight schedule of the experiment, it was decided to use the HATS, with microphones at the same locations, instead of real human subjects. As the pinna molds of the head correspond to an average human pinna, the resultant transfer function would serve as a mean for later equalization of the non-individual binaural reproductions.

B&K's PULSE dual channel FFT analyzer was used to measure the headphone transfer function of the microphone response against the white noise generator. The final response is an average of 5 times, in which the headphones were put on the head anew, in order to accommodate for small changes in the

Left Channel			
Band	1	2	3
$f_c(Hz)$	25	190	2500
Gain (dB)	+10	+3	+4
Q	1.6	3	0.4
Right Channel			
Band	1	2	3
$f_c(Hz)$	25	360	2500
Gain (dB)	+13	+3	+3.8
Q	0.6	2.7	1.2

Table 4.1: The SAE approximated parametric equalizer settings to equalize the Stax headphones at low and midrange frequencies

positioning of the headphones over the ears. The Stax preamplifier volume knob was fixed on one position (at 12 o'clock).

The response of both ears is not identical, but it is very similar to the reported response of the Lambda headphones from [33], shown in figure 4.7. The method described there, involves equalization through Matlab computed IIR filters (Infinite Impulse Response). For the scope of this work, it was decided instead to make use of existing analogue filters and equalize mainly the low and midrange frequencies of interest. Another reason for that is that the correction for the sharp dip above $10kHz$ introduced a 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. SAE 2800 parametric equalizer's resonant filters (see figure 4.9) were employed to correct for the transfer functions. They use a variable center frequency, bandwidth and level of amplification or attenuation. The final settings of the equalizer's bands are given in table 4.1.2. The headphones were equalized only up to a few kHz. It is interesting to note that the headphone response measured on an artificial ear shows a nearly flat curve, apart from a dip at around $4kHz$, designed to counter the ear canal resonance. [2] It indicates that the frequency response measured is mostly due to the HATS outer ear shape and not due to the headphones and most likely not because of the microphone either.

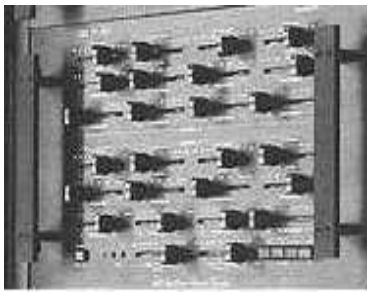


Figure 4.9: SAE 2800 parametric equalizer

4.2 Room Data Acquisition

4.2.1 Reverberation Time Measurements

The setup used for the impulse function measurements in all rooms in shown in figure 4.10.

[2] As was measured later by Brent Kirkwood.

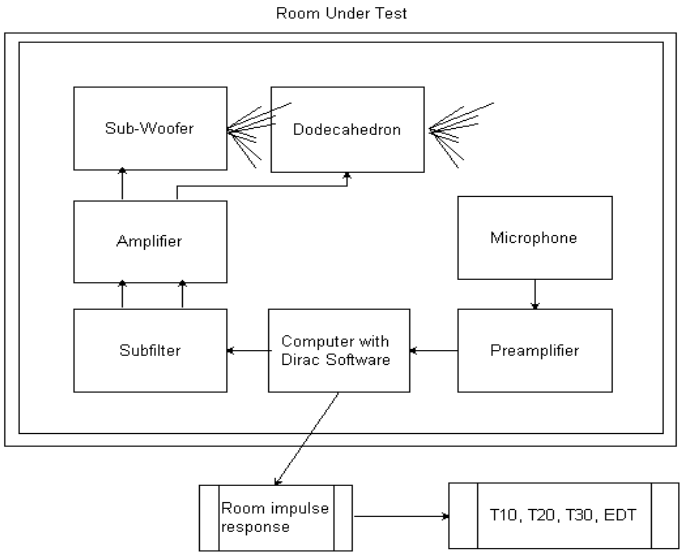


Figure 4.10: Block diagram of room impulse function measurement procedure

Dirac is a software by Acoustics Engineering (AE), which calculates a variety of room acoustics parameters through its impulse response. Version 2.5 was used. The software extracts the impulse response function of the room through recorded predefined signals such as sine sweep and MLS. In that way the impulse response has both high signal to noise ratio and it is faster and more accurate than the classical impulse measurement techniques. In this measurement the impulse function was used to measure the reverberation time through T_{10}, T_{20}, T_{30} and EDT.

Its output was direct to an Amadeus Sub Filter crossover, which split the signal so that one channel in the amplifier (a Labgruppen 300 amplifier) feeds the subwoofer with the lowest frequency content, whereas the other feeds the dodecahedron with the rest. This way the bass frequencies SPL is powerful enough to be picked up by the microphone with a decent signal to noise ratio. The microphone used was a Sennheiser MD 211N.

4.2.2 Dirac's Filter Testing in the Anechoic Chamber

The above setup was built in the large anechoic chamber in the Acoustic Department, DTU, building 354. The anechoic chamber has zero reverberation time at frequencies above 50 Hz. It was repeated a few times, but only at one position of source and receiver in the room. The idea behind the measurement was to use the same setup for the reverberation time measurements to be performed in the small rooms later, and to obtain the lower threshold for valid RT figures, which the measurement system can provide reliably. Any figure different than zero seconds signifies the inherent property of Dirac's filters and have no real meaning as RT of the room. This is necessary since some rooms have very short reverberation times, also at very low frequencies.

The results using various settings are shown in figure 4.11. The measurement that had only 11 averages with $5.6s$ MLS period, seems more stable than the other and was compared with the rest of the rooms. It is also the same one that was used throughout the rest of the measurements, only with 10 averages. The same thresholds are drawn next to the rest of the room RT curves in figures 4.21, 4.22 and 4.23 below. It can be seen that the T_{20} and T_{30} room curves never intersect the system threshold, apart from the lowest $25Hz$ one-third octave band for the hearing protector testing room. It intersects the two DR room curves The EDT system curve intersects the DR talk studio and the DR control room as low as

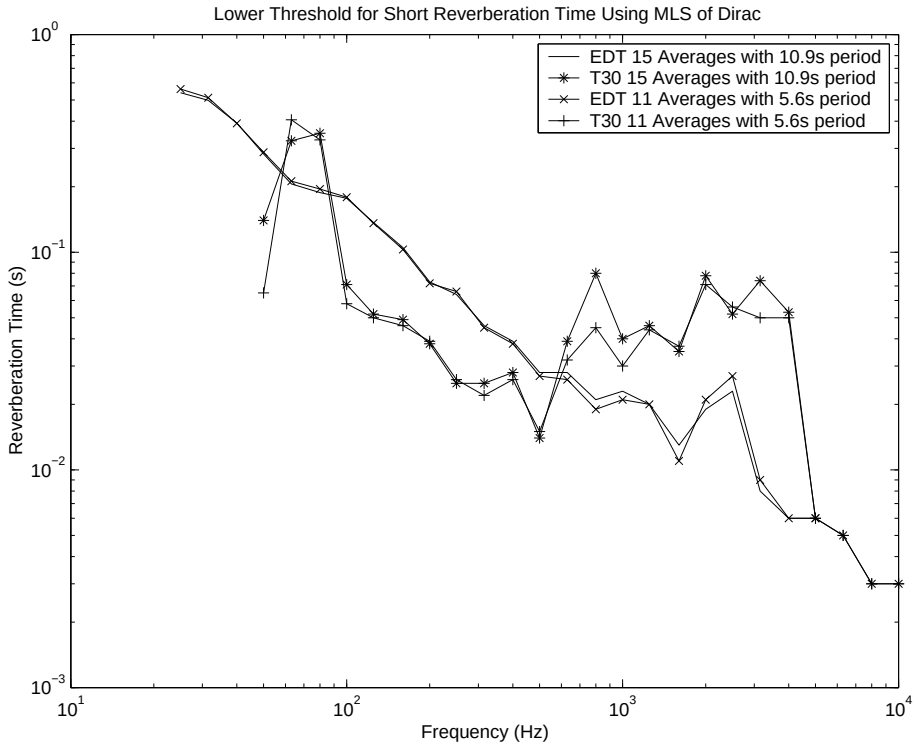


Figure 4.11: Dirac system lowest threshold of short reverberation times measurements

$40Hz$. These bands do have some reverberation in the anechoic chamber, yet probably not as high as the measurements indicated. T_{30} shows a discontinuity at $63Hz$ and $80Hz$, which looks very much like an error. Indeed, the impulse response to noise ratio (INR) of this measurement was very low at this frequencies and consequently the measurement accuracy suffers (see section 4.3.8 and figure 4.24 below). All this comes to show, in agreement with the uncertainty in measurement estimations that were give in section 3.1.3, that the very short RT measurements at the lowest bands should be used with added caution.

4.3 Recordings

The same setup was used for all rooms. Six measurement points (3 independent receiver positions for 2 source positions) were averaged for the final estimations of the room RT. Only in the library 8 positions were measured. In each measurement a $5.46s$ MLS cycle was set to run 10 times. The measurement follows the ISO standard described in [49].

The recording procedure was identical for all 7 rooms measured. It is described in detail only for the first room. The description also mentions the auxiliary procedures that were performed. The rooms were chosen so they represent a wide array of room acoustics, ranging from the well-designed acoustical environments to common rooms that are known to be "bad" rooms.

Note: in all room drawings, except the talk studio, the outer room measurements (floating floors, suspended ceilings and double walls) are not to scale. Also, the lecture and meeting rooms and the library are all constructed using standard Danish floating floors, which are not drawn in the figures.

4.3.1 Pilot Recording - Meeting Room 112, DTU, Ørsted, Building 352

After testing all the equipment available and obtaining the program material, a pilot measurement was performed. It took place in the meeting room (Mødelokale) in Building 352, at the Acoustic Laboratory, DTU.

The following steps were taken:

- Impulse function measurements, using Dirac with MLS. The sequence was $5.46s$ long, well over the maximum expected RT of around $1.5s$ at the low frequencies.
- Noise level measurement using B&K sound level pressure meters Type 2215, in octave bands. All noise levels are summarized in table 4.3 below.
- The room dimensions were measured. A rough plot of the room was drawn, in which the construction materials seen from inside the room were noted, as well as the recording positions. Tables 4.2 and 4.4 below summarize the room dimensions and materials where known.
- Setting the speaker and HATS positions and then the recording levels for all programs. The convenience of having CD that contains all the program material allowed one setting of the recording level for the entire CD. The recording level was set to to avoid clipping in some of the programs. It resulted in a slightly decreased dynamic range for some of the samples.
- Recording - each position of source and receiver was recorded as a separate song on the HDR.

The sound insulation of the room is limited and made it exposed to noise events out in the hall and of cars outside, despite closing the door and windows. Therefore, the level of the programs played had to be rather high, to obtain a reasonable signal to noise ratio. It was around 80 dB(A) SPL.

Surprisingly, the recordings were rather well localized, even without equalization and no longer seemed to be located inside the head. One trick which was employed once was to slightly rotate the HATS, so it is off-axis relative to the speaker. As the reported errors seem to decrease out of the median plane 4.5,

it seemed reasonable to expect better localization achieved in this way, for the price of having a shifted image, louder in one ear than the other. In all the succeeding room recordings the source-receiver were never aligned on-axis. The exact angle on incidence was not recorded and it varied between recordings.

Listening to the speech recordings they clearly sounded unnatural and could not be mistaken for a real speaker, for its bassy and broad image quality. They indeed sound like a loudspeaker played speech, perhaps produced by a loud TV or radio in a domestic environment.

The room is completely bare of any furniture, beside a table in its center, surrounded by chairs.

Four recordings were performed. HATS was always seated on a chair and the loudspeaker was either placed on a stand or on the floor. Their position is shown on the room figure 4.12.

The loudspeaker stand mentioned throughout the recording descriptions is of about $1m$ height.

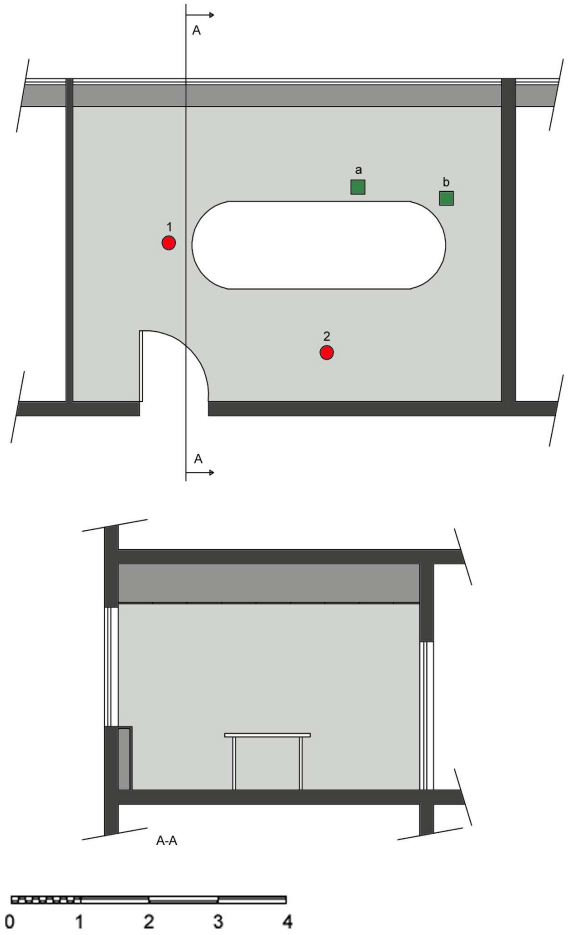


Figure 4.12: Meeting Room 112, Building 352, Ørested, DTU. The room contains a big table surrounded by chairs. The numbered red circles show the source positions and the lettered green squares mark the receiver's. The 4 recorded source-receiver pairs were: 1) 1-a, the speaker was on its stand, the HATS was ox-axis seated on a chair. 2) 1-a, the same, but the HATS is turned 20^0 to the right. 3) 2-b, the speaker is on its stand, the HATS on a chair. 4) 2-b, the same, but the speaker is on the floor.

4.3.2 Library, DTU, 352

The library room has a regular rectangular shape, but its walls are covered with books for the most part. The windows are exposed and so are the upper parts of the brick walls, which are not covered with books. The room is partitioned by book cases, which form a sitting area next to the door, an aisle from which more small aisles lead between the book cases.

It was surmised that the books will absorb much of the sound energy in the room and decrease the reverberation time. That is true for high and mid frequencies, but the books are rather transparent at low frequencies, creating a marked imbalance in the reverberation time curve.

Three recordings were performed. They are showed on the library diagram, in figure 4.13.

The background noise level is higher than in the meeting room, even though it was recorded during the weekend, with less external outside.

4.3.3 Lecture Room, 019, DTU 352

The lecture room has an identical volume as the library, only without the books. All the walls are exposed and there are wooden desks and chairs all over the room, which are highly reflective. It was measured empty, and it is likely that the impulse function with fully occupied classroom is very different. A flutter echo is clearly audible with a hand clap produced by the reflective parallel walls. The RT imbalance, here too, is noticeable and in the played back program the bass was very dominant.

Three recordings were taken, which are showed on the library diagram, in figure 4.14.

The background noise level was the same as in the library.

* * *

At this stage, the three sets of recordings were transferred to the computer, separated into individual tracks and some interim conclusions were drawn, regarding the recording fidelity, binaural replication and audible variations between the rooms.

The two rooms in the Danish Radio are acoustically designed for professional use and fall into the received "good rooms" acoustically, with relatively flat RT curve, irregular shapes to avoid strong standing waves etc.

4.3.4 Control Room, Studio 3, Danish Radio, Fredriksberg, Copenhagen

The control room for Studio 3 is used for recording, producing and mixing of music ensembles was designed accordingly. The RT is relatively low, around $0.3s$ throughout the frequency range. The room is built with double constructions and so it is insulated to a high degree from transmitted sounds through the walls, doors and windows. The shape of the room is irregular and its ceiling height varies within the room. A big section of the room is used for the large 48-channel mixing console. In front of it there are many pairs of loudspeakers, in addition to a pair that is flush-mounted in the wall. Also, there is a large desk behind the console, which also accommodates much equipment.

Three recordings were taken. The speaker was always placed on a shelf just behind the mixing console. In this area of the room there are a few sets of speakers and therefore the location of the speaker is likely in the room. Also it was chosen so that no objects stood in the way between the source and receivers. The recording positions are shown on on the room blueprint in figure 4.16.

The receivers' locations included: the HATS is seated at the rear desk, the HATS is seated at the center of the mixing console and the HATS is seated on the sofa in one of room niches, close to the wall.

4.3.5 Talk Studio 8, Danish Radio, Fredriksberg, Copenhagen

Talk Studio 8 is smaller than the control room, yet exhibits a rather dead acoustics too, due to massive treatment to its walls. Its shape is irregular and the ceiling is sloped. The room is built in double

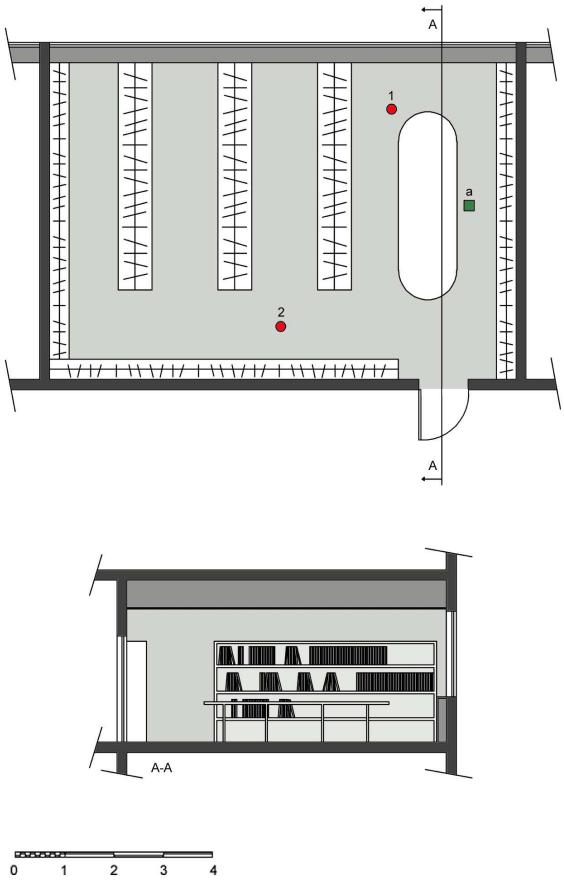


Figure 4.13: Library, Building 352, Ørested, DTU. The numbered red circles show the source positions and the lettered green squares mark the receiver's. The speaker was always on its stand. The HATS was always seated on the same chair in the sitting area, only slightly rotated each time. The 3 recorded source-receiver pairs were: 1) 2-a, the speaker was on its stand. 2) 1-a. 3) 1-a, again, but the HATS was rotated to be more on-axis than before.

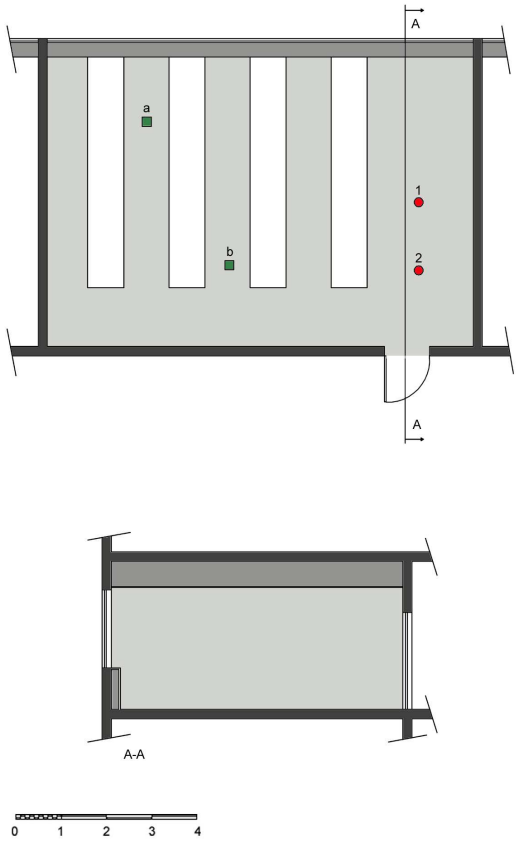


Figure 4.14: Lecture Room 019, Building 352, Ørested, DTU. The numbered red circles show the source positions and the lettered green squares mark the receiver's. The 3 recorded source-receiver pairs were: 1) 1-b, the speaker was on its stand, the HATS seated behind a desk. 2) 1-b, the same, but the speaker was on the floor. 3) 2-a, the speaker was on the stand and the HATS was placed on a desk.

Figure 4.15: Control room of Studio 3 in the Danish Radio, Fredriksberg, Copenhagen

construction to insulate it from the surrounding noise. It is used for talk show broadcasts. It has a desk in one side of the room, close to the door, with all the equipment for the DJ/anchor, plus a few additional microphones for interviewees around the desk (see figure 4.18). At the other end of the room there are a couple of couches with a table, forming a small sitting area. The blueprint of the unfurnished room is shown in Figure 4.17.

In all of the recordings the speaker was placed on the desk, closer to the room center, and rotated to face the general direction of the receiver. The three recording positions were: 1) the HATS sat on a chair at the DJ's position 2) the HATS was seated on a couch close to the wall 3) the HATS was seated on a couch a bit farther from the wall, close to the control room door.

* * *

Finally, two more rooms were measured in Building 354, Ørsted, DTU, to complete the RT curve spread of the room selection.

4.3.6 Hearing Protector Testing Room, DTU, Ørsted, Building 354

This room is used for standardized hearing protectors measurements. It is rather small in volume and it is built from gypsum boards and mineral wool behind them. The room is a floating structure (the room is not connected directly to the rest of the building, but rather floats in another room, insulated). The room exhibits a high average reverberation time for its size and is not meant for uses such as listening room applications. It was designed especially to have a long RT for its size, but with no specific curve shape. The room was empty at the time of the measurements. Figure 4.19 shows a diagram of the room with the corresponding recording positions in it.

4.3.7 IEC Standardized Listening Room, DTU, Ørsted, Building 354

This room complies with the IEC 268-13 standard for Listening Rooms. It is designed to have a flat RT over frequency. At low frequencies the room is more reflective and 4 Helmholtz resonator absorbers were added, which are effective at frequency bands of $125Hz$ and lower [50]. There were no big objects in the room.

Four recordings were done, which are shown on the room drawing in figure 4.20.

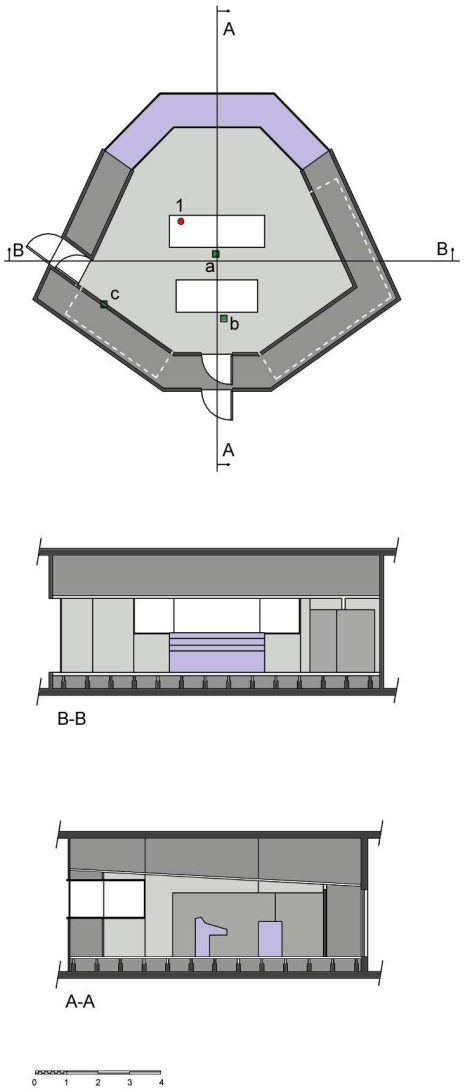
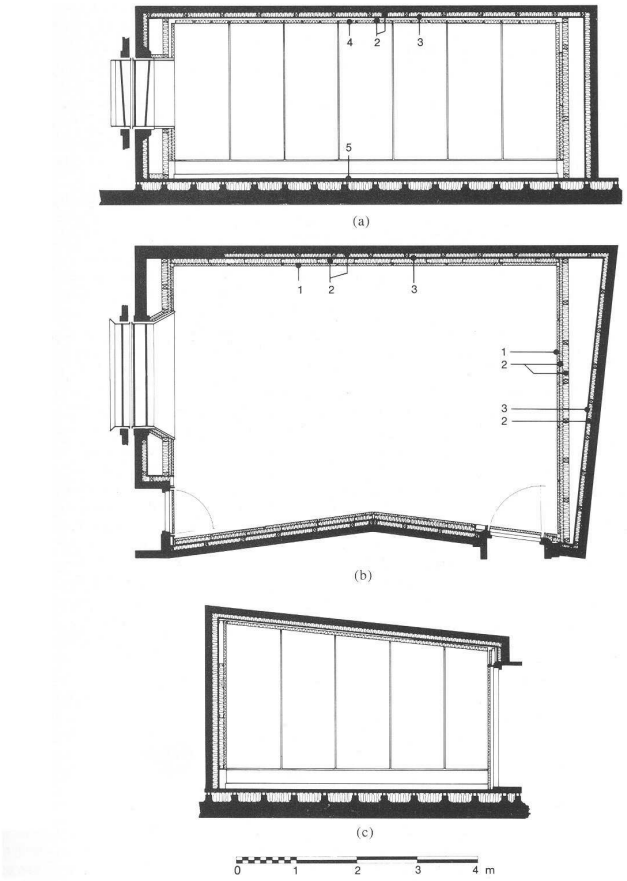


Figure 4.16: Control room for Studio 3 in The Danish Radio, Fredriksberg, Copenhagen. The dashed areas denote niches with lowered ceilings in the room. The numbered red circles show the source positions and the lettered green squares mark the receiver's. The source was placed on the same shelf in all recordings. The 3 recorded source-receiver pairs were: 1) 1-b, the HATS seated on a stool behind the rear desk. 2) 1-c, the HATS was seated on the sofa behind in the niche. 3) 1-a, the HATS was seated on a chair behind the mixing console.



Plan and sections of DR studio 8. 1—40-mm glass wool covered with fabric mounted in steel frames; 2—mineral wool, various thicknesses; 3—13-mm gypsum board; 4—perforated standard metal cassette ceiling; 5—carpet on concrete floor. (a) Longitudinal section. (b) Plan. (c) Cross section.

Figure 4.17: Studio 8 blueprint with a description of the building materials (reprinted from [1]).

Figure 4.18: Talk Studio 8 in the Danish Radio, Fredriksberg, Copenhagen. The photo shows the broadcasting desk in the room and looks at what is the left wall on the blueprints in figure 4.17 (a) and (b).

Room	Length (cm)	Width (cm)	Height (cm)	Volume (m^3)
Meeting	623	418+40	304	84.5
Lecture	945	627+40	301	186.3
Library	945	627+40	301	186.3
Control	N/A			100-110
Talk Studio	N/A			80-85
Hear Prot. Test	413	275	235	26.6
IEC Listening	750	472	275	97.3

Table 4.2: Room principle dimensions. Please note that the volumes of both the control room and the talk studio are only estimated.

4.3.8 Summary

All the 7 room RT and EDT curves are summarized in the following graphs, along with the system threshold (figures 4.21, 4.22 and 4.23. The Dirac software provides another parameter estimation. The impulse response to noise ratio (INR) is the logarithmic ratio of the decay curve to the background noise picked up by the system. The standard [49] requires a minimum of $45dB$ above the background noise level for T_{30} measurements and $35dB$ for T_{20}. However the Dirac manufacturer sets $35dB$ INR as a good measure in practice. Figure 4.24 shows the INR for all measurements. The classroom is the only room in where the INR is lower than 40dB below $63Hz$. Otherwise all rooms show a high INR, above $50Hz$, which is the lowest frequency band used for the analysis later. The anechoic chamber measurement was also susceptible to a lot of noise, mainly because the of the limited sound reflection in the large chamber, which requires much higher output level from the amp to properly excite the microphone.

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 Prot. Test	15	10.3(N)
IEC Listening	15	9.3(N)

Table 4.3: Room noise in NR and RC ratings. (N) stands for Neutral noise spectrum and (R) for Rumbly spectrum.

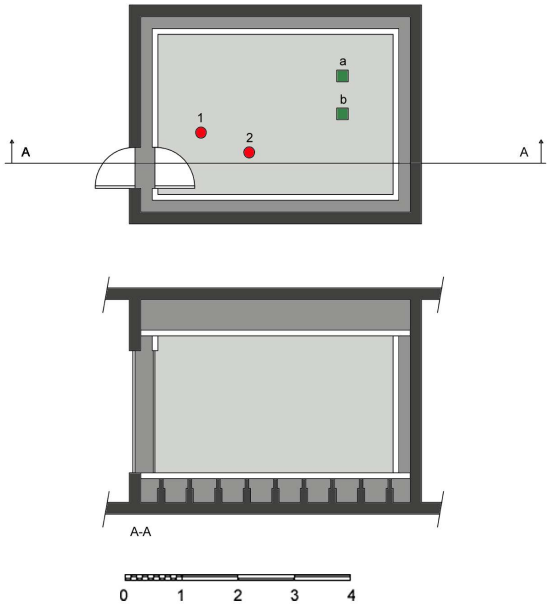


Figure 4.19: Hearing protector testing room in Building 354, Ørested, DTU. The numbered red circles show the source positions and the lettered green squares mark the receiver's. The receiver was seated on the same chair in all the recordings. The 3 recorded source-receiver pairs were: 1) 1-b, the speaker was on the floor. 2) 1-a, the speaker was on the stand. 3) 2-b, the speaker was on the stand.

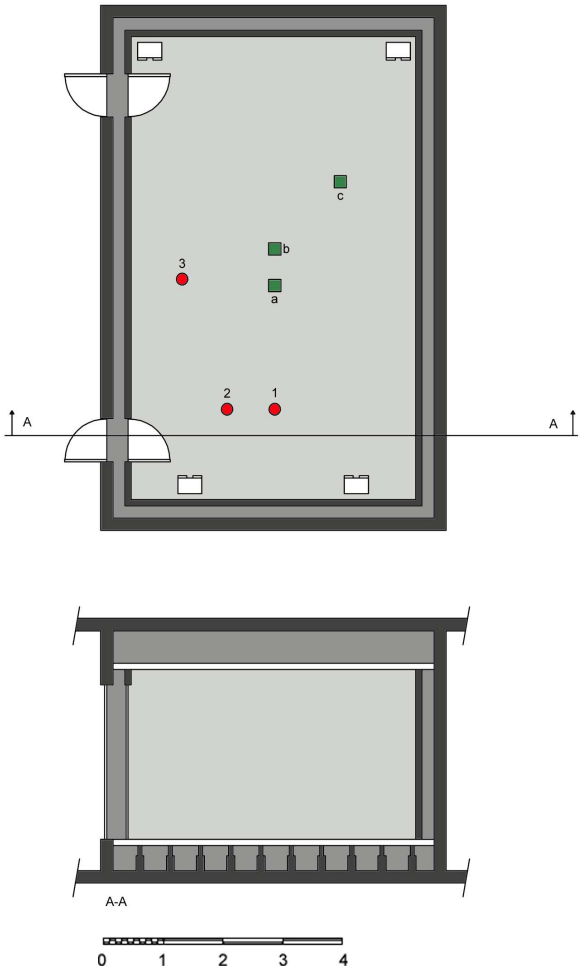


Figure 4.20: IEC Listening Room in Building 354, Ørested, DTU. The numbered red circles show the source positions and the lettered green squares mark the receiver's. The 4 recorded source-receiver pairs were: 1) 1-a, the HATS seated on a couch, the speaker on the floor. 2) 1-b, the HATS is still on the couch, but turn to the speaker's new position 3) 2-b, the speaker is put on its stand and the HATS is seated on a normal chair 4) 3-c, as 3).

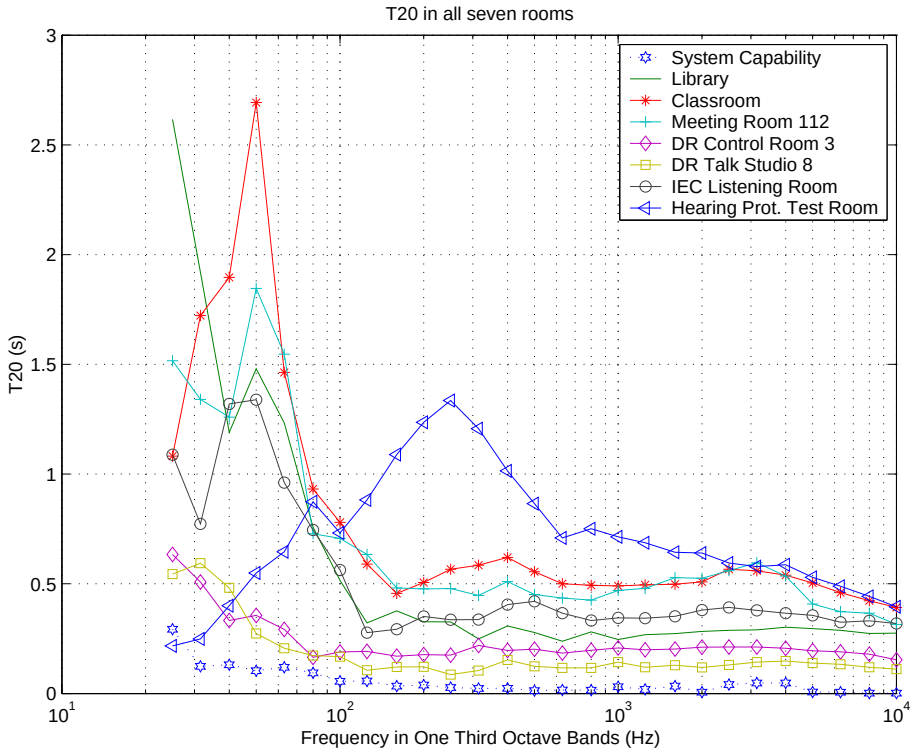


Figure 4.21: The measured average $T20$ in all the rooms in the experiment

Room	Floor	Walls	Ceiling
Meeting	$22mm$ parquet on joists with a cavity of around $10mm$ air	Brick walls with one gypsum wall;windows form another wall; They start from about 90cm, where a hollow box extends the outside wall into the room.	Suspended ceiling with mineral wool layer and air cavity of about $40cm$ below the concrete roof
Lecture	same as in the meeting room, without the gypsum wall		
Library	same as in the meeting room, without the gypsum wall		
IEC Listening	see the meeting room above for the floor description; It is also carpeted.	Uncertain	
Control	uncertain materials and dimensions		
Talk Studio	carpet on concrete floor	40mm mineral wool in steel frames cover with fabric with a 13mm gypsum board behind	perforated standard metal cassette ceiling
Hear Prot. Test	Suspended floor lying on mineral wool springs with no joists	three gypsum layers with mineral and air cavity behind	

Table 4.4: Room Materials

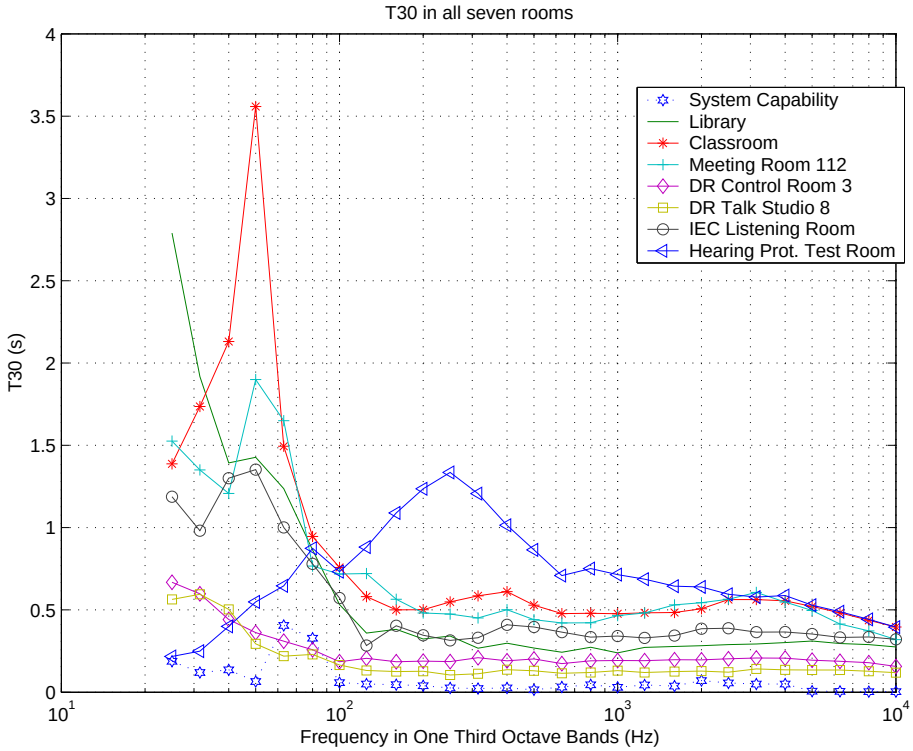


Figure 4.22: The measured average $T30$ in all the rooms in the experiment

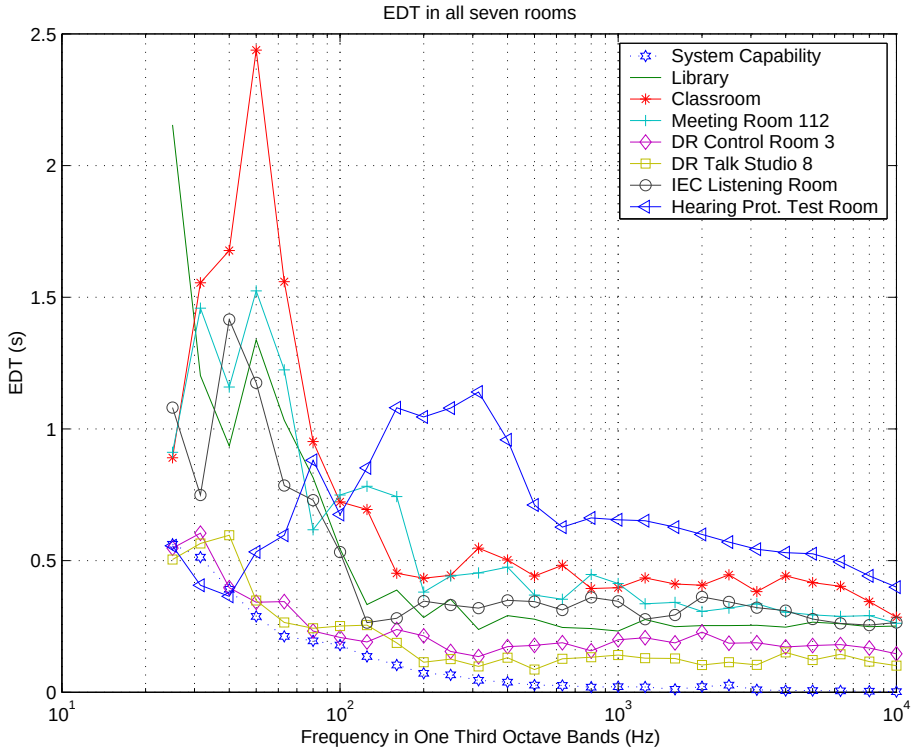


Figure 4.23: The measured average EDT in all the rooms in the experiment

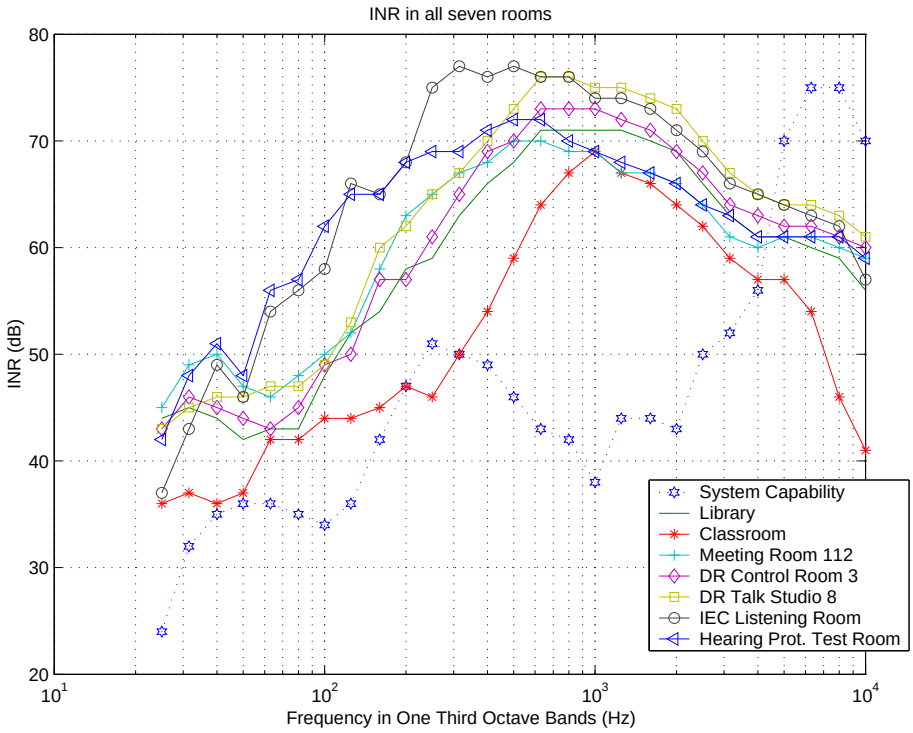


Figure 4.24: The measured average INR in all the rooms in the experiment

Chapter 5

Listening Tests

This chapter describes the listening test constructions and presentation scheme that was used. The following topics are covered:

- sample library construction
- listening test overview
- Experiment A - sound quality, boominess and boxiness rating test
- Experiment B - room matching test
- Experiment C - room matching test with different programs
- general presentation details

5.1 Sample Library Construction

With 230 tracks recorded in total at 23 different positions in different 7 rooms, obviously not all recordings can be used, as for time and budget constraints. All tracks were transferred from the HDR, where they were stored in song format, to a PC for a library formation according to track name, room and position number in the room. Adobe's Audition 1.0 software was used to trim the tracks and normalize them to an even output level where was necessary.

Several points were noted regarding specific tracks.

White and pink noise all sounded different from one another and, in turn, from the original white noise. However, no "anchors" could be found by close listening to them, in regard to apparent coloration as was expected. Similar recordings in the past, which were taken with reel-to-reel tapes, revealed coloration more easily when doubling the playback speed. The expected effect is of tonal components being brought out of the uniform white noise. The same manipulation was attempted here too, by doubling the white noise playback speed, but to no avail, as no interpretable components came out of the overall noise. Therefore, white and pink noise recordings were not used altogether in any test.

Recordings of Track 5 (Gismonti) tended to be distorted, due to its over-amplified level on the playback CD. Hence, track 5 was not used for any test.

The use of track 6, a medieval piece by Machaut, was kept to a minimum. The continuous character of the music, along with an omnipresent legato and the original long RT of the recording make any judgment more difficult than in other pieces.

5.2 Listening Test Overview

The framework decided upon in advance was a rating scale test of sound quality preference listening test with the existing material, using test subjects of various backgrounds. In addition, the test subjects would be asked to rate the recording "boominess" and "boxiness", in attempt to later correlate them to longer RT at low frequencies and to perceived coloration, respectively. An objective room parameter that possibly correlates with boominess is the room bass ratio (see 2.3). Since this is essentially a preference test, there is a risk of 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 that they may have to deal with.

In light of all these, it seemed appropriate to venture yet another test, which makes use of the same samples. The basic idea evolved from a number of sources. As was mentioned before, recordings in each room were done in 3 or 4 positions of source and receiver, in order to get a more comprehensive picture of each room in the preference tests. Although those 3-4 recordings of each room sometimes sound markedly different than one another, they all share some common acoustical characteristics of the room. Can a listener always tell between samples from one room and samples from another, even though the recording positions entail quite a different sound of each sample? This question is not trivial when the two rooms are acoustically similar, and therefore it seemed worthy of closer examination.

The following three listening tests were introduced. All of them were automated and computerized using Matlab codes. The main reason for using a completely automated test, was the staggering amount of samples that have to be played back in a random order, along with matching bookkeeping of the subject's answers. The codes are given in Appendix B.

The playback setup was also identical, and used the playback chain that was described in the previous chapter. Headphone playback permitted having the test in a normal (non-listening) 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 on any noise related problems to solve the tests.

The subjects were given no time constraints for solving the tests and normal running times varied between an hour and a half to two and a half hours.

Subjects were instructed to take breaks between the two parts of each experiment and between the experiments, in order to avoid fatigue.

Due to the overall multitude of samples used, the samples did not have the exact volume level. Sometimes the difference was very apparent and subjects were instructed to adjust the volume, when necessary, using the Stax preamplifier's volume knob. Nevertheless, in most cases, the listeners did not need to change the presentation levels.

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

Finally, 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 4.3.1) - when making decisions.

5.3 Experiment A - 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 have to evaluate. The subjects are requested to note their impression of the overall sound quality, boominess and boxiness of the samples according to the following written guidelines:

"Overall Sound Quality" - is 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 /

Room	Source-Receiver Position
DR Studio 3 Control Room	2
DR Talk Studio 8	1
Hearing Protector Room	1
IEC Listening Room	1
Lecture Room	1
Library	1
Meeting Room	3

Table 5.1: Source-Receiver Position for rooms in experiment A. Refer to section 4.3 for details about the exact locations in the rooms.

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 in in which the bass lacks definition and 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 times when you talk inside a closet (if you hid there as a child), as typical illustrations.

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

For familiarization with the scales used and the degree of variation over samples, three examples were given of another recording (Track 8, by Jimi Tenor) in three very different rooms: the talk studio, lecture room and hearing protector testing room.

Samples can be played back as many times as the subject wishes, before the three ratings are given one after the other.

The presentation order is randomized using a Latin square scheme and the 14 samples are presented again in a different order after a short break.

The male speech recordings (Track 3) were chosen for speech evaluation, for their higher spectral content, compared to female speech.

Track 7, by Herbert, 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 both on the detail resolution, the effects of reverberation time on single events and on continuous signals at the same time. All in all, its instrumentation is simple enough, in order for a listener not to lose track in a sea of sound. See section 4.1.1 for more details about the programs.

The representative positions in the room, chosen for the two programs are given table 5.1. The numbering refers to the room description in section 4.3.

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

For boominess rating the scale was:

Question	Room 1 (2 Positions)	Room 2 (2 Positions)	Program Name
Example A	DR Talk Studio (3,2)	Hearing Protector Room (1,3)	Female
Example B	IEC Listening Room (2,4)	Library (1,3)	The The
1	Lecture Room (1,2)	Library (1,3)	Jimi
2	DR Control Room (1,2)	DR Talk Studio (1,3)	Mingus
3	Meeting Room (2,4)	Library (1,3)	Herbert
4	DR Talk Studio (2,3)	IEC Listening Room (1,2)	Male
5	Meeting Room (1,4)	DR Control Room (1,3)	Jimi
6	DR Talk Studio (1,2)	Library (1,3)	Male
7	Meeting Room (2,3)	Hearing Protector Room (1,3)	Mingus
8	DR Conrtol Room (1,3)	IEC Listening Room (3,4)	Herbert
9	Lecture Room (1,3)	Hearing Protector Room (1,2)	Male
10	DR Control Room (2,3)	Library (1,3)	Mingus
11	Meeting Room (1,3)	Lecture Room (1,2)	Herbert
12	Library (1,2)	IEC Listening Room (1,3)	Jimi

Table 5.2: Room pairs - test items for Experiment B. The number in parentheses indicates the recording positions used, corresponding to the room data numbering in section 4.3.

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 confusions.

The boxiness scale is unipolar:

1. Unnoticeable
2. Barely Audible
3. Distinct Yet Not Dominant
4. Dominant
5. Very Dominant

5.4 Experiment B - Room Matching Test

This experiment was specially designed to test whether subjects are able to discriminate and match a pair of recordings, which was recorded in the same room, but at different source and receiver positions. Four programs were used for the test: male speech, Herbert (track 7), Jimi Tenor (track 8) and Mingus (track 9; see section 4.1.1 for more program material details). The 7 rooms were divided into pairs. Out of 21 possible pairs of rooms, only 12 were tested, in varying difficulty levels. Difficult pairs were considered to be of two rooms with a very similar RT curve (see 4.3). Each program was used for 3 pairs. In table 5.2 the 12 pair description is given in detail.

The test program has an interface which is shown in figure 5.1. The user can play each of the samples and toggle between them by pressing on a letter button and determine which of the samples B, C or D was recorded in the same room as sample A. Pressing a button plays a sample from its beginning. Sometimes it may be slightly easier for a subject to match the other pair, which was recorded also in one room. Two examples were presented for a familiarization round (see table). Other programs were used, which were not used later in the test itself. The subject was informed whether his/her choice was correct and in case

it was incorrect, another chance was given to listen to the correct pair. The first example was very easy and the second more difficult. Two rounds of the 12 questions were thereafter presented with a break in between. The questions were randomized using a Latin square scheme and also the assignment of each sample to be either A, B, C or D was completely random. After each round the subject was informed of the number of successfully recognized pairs out of 12 questions. In the later exported output file a correctly recognized pair is marked by 1 and an incorrect answer is 0.

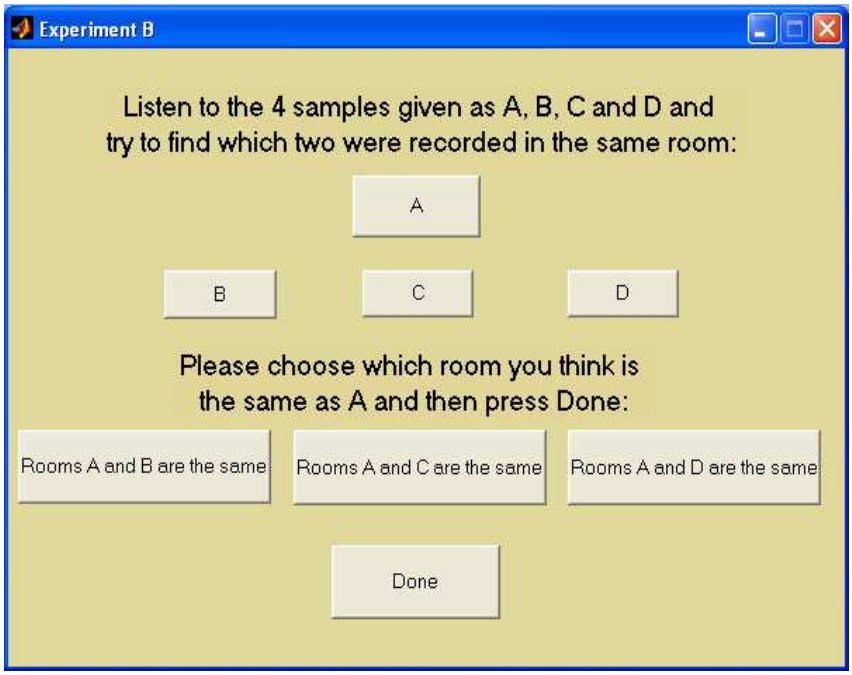


Figure 5.1: The interface for experiments B and C.

5.5 Experiment C - Room Matching Test with Different Programs

The last experiment is identical in procedures to Experiment B, but is much more difficult. This time the four samples are four different programs, still divided into two pairs. Two rounds of 8 questions were presented with a break in between. The 8 pairs and their preceding examples are give in table 5.3. The subjects were not given any special instructions or hints beyond what was already known from the previous test.

5.6 General Presentation Details

5.6.1 Test Subjects

A total of 18 subjects were tested on the three experiments, including one pilot test subject. After the initial pilot tests only slight modifications were done, and so the pilot subject's data was used in the final analysis.

Question	Sample 1	Sample 2	Sample 3	Sample 4
Example A	Talk 3 (Female)	Talk 2 (Jimi)	HP Room 1 (Machaut)	HP Room 3 (The The)
Example B	IEC 2 (The The)	IEC 3 (Female)	Library 1 (Machaut)	Library 3 (Jimi)
1	Meeting 2 (Jimi)	Meeting 4 (Male)	Lecture 3 (Mingus)	Lecture 1 (Herbert)
2	Control 2 (Herbert)	Control 1 (Jimi)	Talk 2 (Mingus)	Talk 3 (Male)
3	Meeting 3 (Jimi)	Meeting 1 (Mingus)	Library 2 (Herbert)	Library 1 (Male)
4	Talk 1 (Herbert)	Talk 3 (Jimi)	IEC 3 (Male)	IEC 4 (Mingus)
5	Meeting 1 (Male)	Meeting 4 (Mingus)	Talk 1 (Jimi)	Talk 3 (Herbert)
6	Library 2 (Male)	Library 1 (Jimi)	Talk 2 (Herbert)	Talk 3 (Mingus)
7	Lecture 3 (Jimi)	Lecture 2 (Male)	HP 3 (Herbert)	HP 2 (Mingus)
8	Control 3 (Male)	Control 2 (Jimi)	IEC 1 (Herbert)	IEC 3 (Mingus)

Table 5.3: Room pairs - test items for Experiment C. The number indicates the recording position, corresponding to the room data numbering in section 4.3. The program name is given in parentheses.

Test subjects were between 20 to 40 years old and all were tested for normal hearing either recently or before the tests, according to the ISO 389 standard for hearing threshold measurements.

The test subjects included 14 males and 4 females, of whom 13 are students or staff at the acoustic department and 5 are completely untrained subjects (as far as room acoustics is concerned).

5.6.2 Presentation Order

The listening tests were presented in the order of Experiment B, Experiment A and Experiment C. The reasoning for that was twofold. First, by having Experiment B (room pair matching test) in the beginning, subject will get superficially familiarized with the range of recordings to be used later in the rating test. Second, by not having the same task repeated immediately (Experiments B and C), the subject alertness and interest may be more easily maintained.

For the sake of simplicity the discussion below will present the results according the original design of A, B and C.

5.6.3 Presentation Level

Subjects were instructed to listen to the samples at a comfortable level and thus, if samples are confusingly different in level, to compensate for that using the volume knob of the Stax preamplifier. Still, most subjects tended to leave it at more or less a constant 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 $1000Hz$ using a sound level calibrator B&K Type 4230). The A-weighted level of samples varied between $70dB$ to 80, depending on the program material, the particular recording and the individual volume knob setting.

There was a concern that variable presentation level will unwantedly influence the bass perception of subject, due to the compression of bass dynamics in loudness perception. Equal contour loudness levels are strongly dependent on the sound pressure level (see figure 5.2). At lower levels the bass is perceived to be weaker, compared to higher frequencies. The higher the SPL goes, the more the bass evens out. At levels corresponding to our presentation levels (around $50 - 60dB$ SPL at $1kHz$), the differences between curves is not as pronounced as for lower SPL's. However, the curves were measured for pure tones, and the application to complex signals is not straightforward as that.

Two reasons for not fixing the presentation levels were argued. 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 something which is, in real life, optional in domestic situations and so, a subject, who is to give a critical opinion should do so under his preferable conditions.

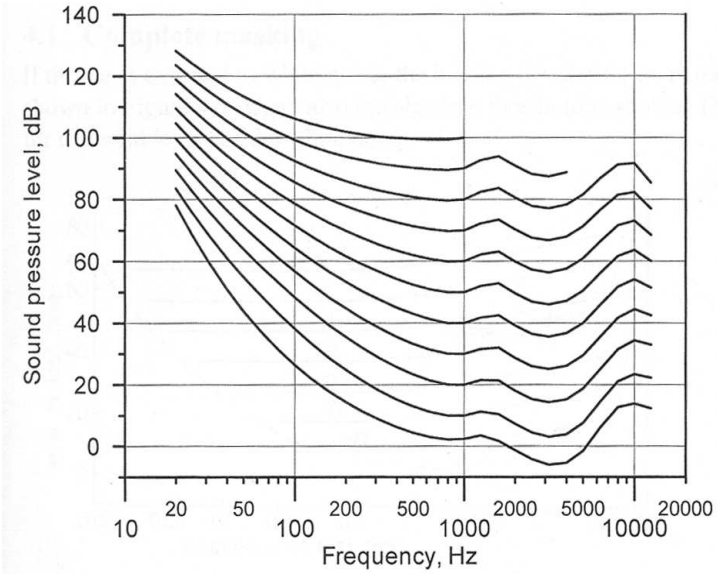


Figure 5.2: Equal loudness level contours loudness levels for pure tones. From ISO-226, 2002.

The main concern in the variable presentation level was for the boominess ratings. That question will be examined later in the analysis.

5.6.4 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 that it will convey a realistic sensation to listeners.

From haphazard sampling of listeners, it seems that all subjects at times experienced 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, but it seems reasonable to assume that where the original source was located off-axis, replication is more realistic. That effect of unbalanced recordings was at time confusing for subjects.

Chapter 6

Test Results and Analysis - Experiment A

The results of the three experiments that were described in the last chapter are presented in this chapter and in the next, along with their statistical and acoustical analysis.

All statistical analyses were performed used Matlab statistical toolbox.

6.1 Task Fulfillment

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 over-analyze 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, for 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, due to his dislike for it. His remark had later eased the verification of the halo error correction described below.

At least 2 subjects reported that their rating of overall sound quality of the speech samples was directly related to their impression from the speech intelligibility, which was always rather high. Their ratings were used notwithstanding, as their interpretation is seen as a legitimate element of sound quality.

Some subjects complained about initial difficulty to rate boominess, while others had difficulties with boxiness.

A few subjects reported a mistyping error, which yielded an unintentional rating. Unfortunately, these errors could not be corrected and they contribute to the total error in the measurement.

6.2 The Preparatory Analysis

The rating data from Experiment A is composed of 4 factors: rater, room, program and parameter. The analysis began by separately analyzing each parameter. A further element was the double rating of each room by each rater, due to the use of test and re-test 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 and they are exemplified for the overall sound quality only. The analysis includes the following stages: a comprehensive four-way ANOVA for each parameter including all factors; halo effect error correction; individual linear transformation of ratings to conform for a regular scale with invariable dispersion between raters; derivation of all means per rooms and programs; correlations and connections to room acoustic parameters.

Since the rather lengthy statistical analysis may be too laborious and tedious for some, the reader is welcome to skip the following sections directly to the results, starting from section 6.3.

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

6.2.1 Overall Sound Quality Analysis of Variance

An overview of all the 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 below in the ANOVA[1] summary table (figure 6.1).

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Raters	278.29	17	16.3697	12.5	0
Rooms	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
Raters*Rooms	286.58	102	2.8096	2.14	0
Raters*Program	98.04	17	5.767	4.4	0
Raters*Part No.	15	17	0.8824	0.67	0.8288
Rooms*Program	125.47	6	20.912	15.96	0
Rooms*Part No.	5.57	6	0.9279	0.71	0.6431
Program*Part No.	0.51	1	0.5079	0.39	0.5339
Error	430.95	329	1.3099		
Total	1614.5	503			

Figure 6.1: Four-way ANOVA summary table of the overall sound quality ratings in Experiment A

The only insignificant factor in the experiment is the test part, including second-order interaction terms with it. It means that the average ratings are stable between the two parts, as can be easily double-checked, by calculating the reliability of the test. 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 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 Raters/Program term in not obvious and is also undesired in the test. It points to a halo effect error, since subjects showed biased response or preference to a certain program over the other. Fortunately, this particular error can be corrected for, as is shown below. Before that we can note the statistical rating model, which is used throughout:

$$SQ = \mu_{SQ} + X_{Rater} + X_{Program} + X_{Room} + X_{Room*Rater} + X_{Room*Program} + X'_{Rater*Program} + \epsilon \quad (6.1)$$

Where the total measured rating is composed from the mean rating μ, and the X_i's are the effects of individual raters, programs and room. The two room interaction terms were added from observation at the table. They account for a large fraction of the total variance. They mean that subjects rate different

[1]The condition for ANOVA states that the data has to be normally distributed and homogeneous (in respect to its variance. Only the four-way ANOVA's 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.

rooms differently and that the room rating is not invariable to the program played. $X'_{Rater*Program}$ is the interaction term between raters and programs, which is the halo effect error. In the following correction it appears as X'_{ki}, in convention with Guilford [41]. The rest of the error in the rating is given by ϵ. The boxiness (figure 6.2) and boominess (figure 6.3) models include less terms and so they are simpler:

$$Boxiness = \mu_{box} + X_{Rater} + X_{Room} + X_{Room*Rater} + X_{Room*Program} + X'_{Rater*Program} + \epsilon \quad (6.2)$$

$$Boominess = \mu_{boom} + X_{Rater} + X_{Room} + X_{Room*Rater} + X_{Room*Program} + X'_{Rater*Program} + \epsilon \quad (6.3)$$

Where the effect of the program alone does not play a role in either rating. We will return to these model later.

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Raters	38.912	16	2.432	3.49	0
Rooms	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
Raters*Rooms	210.059	96	2.1881	3.14	0
Raters*Program	37.248	16	2.328	3.34	0
Raters*Part No.	12.139	16	0.7587	1.09	0.364
Rooms*Program	12.992	6	2.1653	3.11	0.0057
Rooms*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			

Figure 6.2: Four-way ANOVA summary table of the boxiness ratings in Experiment A

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Raters	61.161	17	3.5977	4.42	0
Rooms	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
Raters*Rooms	126.381	102	1.239	1.52	0.0031
Raters*Program	43.558	17	2.5622	3.15	0
Raters*Part No.	29.732	17	1.7489	2.15	0.0055
Rooms*Program	17.27	6	2.8783	3.54	0.0021
Rooms*Part No.	15.246	6	2.541	3.12	0.0055
Program*Part No.	0.002	1	0.002	0	0.9607
Error	267.875	329	0.8142		
Total	623.839	503			

Figure 6.3: Four-way ANOVA summary table of the boominess ratings in Experiment A

6.2.2 Correction of the Halo Effect Error

A short glimpse at the mean ratings in respect to the program rated 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. It was explained in a previous section as the Halo Effect, a common error in rating tests. The correction procedure is rather cumbersome. It will be described briefly here and the corrected results for the present results are given and used.

The presence of the halo effect can be inspected using two-way ANOVA on a factorial design matrix, which has one dimension as raters, and the other as programs. The rating in each cell is the mean of the seven rooms for each category. In case that the effect is indeed present, the ANOVA interaction result will be significant, i.e. ratings for the different programs originating from all raters do not represent the

Raters\Programs	Music	Speech	All Prog.	X'_{kl}
1	5.0000	4.1429	4.5714	-0.5238
2	5.0000	4.4286	4.7143	-0.3810
3	4.5714	6.2857	5.4286	0.3333
4	4.2857	5.1429	4.7143	-0.3810
5	4.2857	5.2857	4.7857	-0.3095
6	4.7143	5.8571	5.2857	0.1905
7	3.5714	4.0000	3.7857	-1.3095
8	6.0000	6.7143	6.3571	1.2619
9	5.5714	6.4286	6.0000	0.9048
10	3.0000	4.0000	3.5000	-1.5952
11	5.2857	4.8571	5.0714	-0.0238
12	6.4286	5.7143	6.0714	0.9762
13	5.5714	5.7143	5.6429	0.5476
14	5.0000	6.0000	5.5000	0.4048
15	4.5714	5.8571	5.2143	0.1190
16	4.8571	5.5714	5.2143	0.1190
17	4.8571	5.1429	5.0000	-0.0952
18	4.7143	5.0000	4.8571	-0.2381
All raters	4.8492	5.3413	5.0952	
d_i	0.2460	-0.2460		

Table 6.1: Halo Effect for overall sound quality in part 1 - means of programs derived from ratings by different raters

same population. Once significance is established, the individual means for programs and individuals may be adjusted according to the general mean of all ratings. The procedure follows closely Guilford [41], although in the current experiment there is one additional factor (test part). The analogy was done by substituting ratees for programs and traits for rooms. Ratings for overall sound quality, boxiness or boominess were assumed independent and the procedure was repeated for each one. An example of this process is given for the correction of the sound quality ratings of the first part.

The following ANOVA table shows a significant interaction between rater and program below 1%. Table 6.1 shows the input matrix, with averages over rooms for all raters and programs. Individual rater's means are given in the third column and program's mean in the second to last row. The overall mean in shown in their intersection. The deviations from the overall mean by all other means are given in the last row and column.

```
Two-way parametric ANOVA of part 1's overall sound quality, ignoring
differences between rooms.
Source                SS     df         MS         F         Prob>F
Columns         167.4286     17     9.8487     3.7154     2.9375e-006
Rows             24.1429      1    24.1429     9.1078         0.0029
Interaction      94.5714     17     5.5630     2.0986         0.0080
Error           572.5714    216     2.6508
Total           858.7143    251
```

The cell values SQ_{ikl} are corrected according to their deviations using:

$$SQ'_{ikl} = SQ_{ikl} - X'_{kl} - d_i \tag{6.4}$$

Where SQ'_{ikl} is the corrected value, X'_{kl} is the rater error and d_i is the program deviation. The new values are shown in table 6.2. The new values are adjusted so they inherently have the same program and rater means, which equals the overall mean.

Finally, the difference is found between the cell value and the global mean. It is shown in table 6.3. This difference is the halo error for each subject for one of the two programs. It can now be subtracted from all the original sound quality ratings. For instance, it is seen that subject 3, who reported his bias

Raters\Programs	Music	Speech	All Prog.
1	5.7698	4.4206	5.0952
2	5.6270	4.5635	5.0952
3	4.4841	5.7063	5.0952
4	4.9127	5.2778	5.0952
5	4.8413	5.3492	5.0952
6	4.7698	5.4206	5.0952
7	5.1270	5.0635	5.0952
8	4.9841	5.2063	5.0952
9	4.9127	5.2778	5.0952
10	4.8413	5.3492	5.0952
11	5.5556	4.6349	5.0952
12	5.6984	4.4921	5.0952
13	5.2698	4.9206	5.0952
14	4.8413	5.3492	5.0952
15	4.6984	5.4921	5.0952
16	4.9841	5.2063	5.0952
17	5.1984	4.9921	5.0952
18	5.1984	4.9921	5.0952
All raters	5.0952	5.0952	5.0952

Table 6.2: Halo Effect for overall sound quality in part 1 - means corrected for rater errors X'_{kl} and for program deviation d_i

for the unfavorite music over speech, shows indeed the second biggest error. In contrast, rater 7 shows a remarkably small error, showing an unbiased program rating.

The same process was repeated for: SQ for part 2, boominess for part 2 and boxiness for part 2. The boominess and boxiness ratings for part 1 did not show a significance of less than 5% and were not corrected.

6.2.3 Linear Transformation of The Ratings

The ratings by different individuals naturally have different means and dispersions. So for instance, subject A may center his SQ rating around 5 and rate only between 4-6, whereas another's is centered around 6 and rates all between 4-8. As we are interested with the eventual relative ranking of the rooms, it is helpful to conform all ratings to a standard scale, with one mean and one dispersion. This way ensures that all rater's ratings are weighted equally [41]. The ITU standard suggests a simple transformation to perform this normalization [6]:

$$Z_i = \frac{x_i - \overline{x_i}}{s_i} \cdot s_t + \overline{x_t} \tag{6.5}$$

where: Z_i is the normalized result, x_i is the rating of subject i, $\overline{x_i}$ is the mean rating of subject i, s_i is the subject's standard deviation, s_t is the total standard deviation for all subjects and $\overline{x_t}$ is the total mean.

However, this type of normalization does not correct for the central tendency error that was mentioned earlier. A rater who for any reason made only limited use of the entire breadth of the scale is bound to deliver a limited discrimination of the rated objects, regardless of the normalization, which only shifts its mean and stretches or compresses an individual scale.

6.2.4 ANOVA Revisited

After applying this normalization the four-way ANOVA is repeated (figure 6.4 below. 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 the probability is artificially brought to 1 - no interaction. The normalization of the ratings reduced the overall variance - first, by diminishing the variance over raters

Raters\Programs	Music	Speech	$\sum$
1	0.6746	-0.6746	0.0000
2	0.5317	-0.5317	0.0000
3	-0.6111	0.6111	0.0000
4	-0.1825	0.1825	0.0000
5	-0.2540	0.2540	0.0000
6	-0.3254	0.3254	0.0000
7	0.0317	-0.0317	0.0000
8	-0.1111	0.1111	0.0000
9	-0.1825	0.1825	0.0000
10	-0.2540	0.2540	0.0000
11	0.4603	-0.4603	0.0000
12	0.6032	-0.6032	0.0000
13	0.1746	-0.1746	0.0000
14	-0.2540	0.2540	0.0000
15	-0.3968	0.3968	0.0000
16	-0.1111	0.1111	0.0000
17	0.1032	-0.1032	0.0000
18	0.1032	-0.1032	0.0000
$\sum$	0.0000	0.0000	0.0000

Table 6.3: Halo Effect for overall sound quality in part 1 - contributions of interactions of rater and program: Halo errors X'_{ki}

and then by doing the same for the other two interaction terms with raters. Similar data was obtained from the boxiness and boominess, seen in figures 6.5 and 6.6 respectively.

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Raters	29.135	17	1.7138	13	0
Rooms	335.139	6	55.8565	423.84	0
Program	38.889	1	38.8889	295.09	0
Part No.	0.071	1	0.0714	0.54	0.4621
Raters*Rooms	28.059	102	0.2751	2.09	0
Raters*Program	0.067	17	0.0039	0.03	1
Raters*Part No.	1.823	17	0.1073	0.81	0.6771
Rooms*Program	125.472	6	20.912	158.68	0
Rooms*Part No.	5.567	6	0.9279	7.04	0
Program*Part No.	0.508	1	0.5079	3.85	0.0505
Error	43.358	329	0.1318		
Total	608.088	503			

Figure 6.4: Four-way ANOVA summary table of the overall sound quality ratings in Experiment A after halo effect correction and rating normalization

As of formality the superfluous terms with little contribution to variance or those with unwanted effected can be excluded from the final ANOVA models. An example of a reduced ANOVA table is given for the sound quality in figure 6.7. The reduced SQ, boxiness and boominess models can be rewritten, respectively:

$$SQ = \mu + X_{Rater} + X_{Program} + X_{Room} + X_{Room*Rater} + X_{Room*Program} + \epsilon \tag{6.6}$$

$$Boxiness = \mu_{box} + X_{Room} + X_{Room*Program} + \epsilon \tag{6.7}$$

$$Boominess = \mu_{boom} + X_{Room} + X_{Room*Program} + \epsilon \tag{6.8}$$

In the latter two models the rater effect was omitted, despite testing significant. It is because its very small contribution to the total variance. In the boominess ANOVA table in figure 6.6, it can be seen how

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Raters	1.132	16	0.0707	2.07	0.0095
Rooms	137.513	6	22.9188	671.03	0
Program	2.288	1	2.2878	66.98	0
Part No.	0.254	1	0.2542	7.44	0.0067
Raters*Rooms	5.662	96	0.059	1.73	0.0002
Raters*Program	0.241	16	0.015	0.44	0.9708
Raters*Part No.	0.41	16	0.0256	0.75	0.7408
Rooms*Program	12.992	6	2.1653	63.4	0
Rooms*Part No.	1.966	6	0.3277	9.6	0
Program*Part No.	0.002	1	0.0021	0.06	0.8043
Error	10.588	310	0.0342		
Total	173.047	475			

Figure 6.5: Four-way ANOVA summary table of the boxiness ratings in Experiment A after halo effect correction and rating normalization

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Raters	1.853	17	0.109	3.25	0
Rooms	61.548	6	10.2579	306.13	0
Program	1.05	1	1.0496	31.32	0
Part No.	0.018	1	0.0179	0.53	0.4659
Raters*Rooms	3.66	102	0.0359	1.07	0.3236
Raters*Program	0.478	17	0.0281	0.84	0.6475
Raters*Part No.	0.979	17	0.0576	1.72	0.0381
Rooms*Program	17.27	6	2.8783	85.9	0
Rooms*Part No.	15.246	6	2.541	75.83	0
Program*Part No.	0.002	1	0.002	0.06	0.8079
Error	11.024	329	0.0335		
Total	113.128	503			

Figure 6.6: Four-way ANOVA summary table of the boominess ratings in Experiment A after halo effect correction and rating normalization

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.

The following ranking was derived from all data. It should be noted that as the correction above is linear and was performed with no other types of error corrections, the resultant ranking is unaffected by it. However, since the variance of the means is smaller, the confidence intervals are decreased and the significance of means in respect to each other improves. The ranking is determined by the output plot of the one-way ANOVA Multcompare, which illustrates the 95% confidence interval for each room, in respect to all other rooms. Naturally, the more ratings are joined together, the more relative ranks will be significant. An output plot, for the overall sound quality in both parts is given as an example in figure 6.8.

6.3 Analysis Results

Table 6.4 shows the summary of all the ratings from Experiment A. The data was statistically manipulated to extract the maximum amount of data from it.

6.3.1 Correlation to Room Acoustical Data

After having obtained all the means, a basic examination can reveal whether there are any inter-correlations between the three rated parameters and correlations between them and the room acoustical parameters. 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

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Raters	29.135	17	1.7138	12.37	0
Rooms	335.139	6	55.8565	403.21	0
Program	38.889	1	38.8889	280.73	0
Raters*Rooms	28.059	102	0.2751	1.99	0
Rooms*Program	125.472	6	20.912	150.96	0
Error	51.394	371	0.1385		
Total	608.088	503			

Figure 6.7: Four-way ANOVA summary table of the overall sound quality ratings in Experiment A after halo effect correction and rating normalization

Figure 6.8: Four-way ANOVA pair significance test between sound quality of different rooms. The information is summarized in table 6.4. Room numbering key: 1 - DR Studio 3 Control Room, 2 - Talk Studio 8 in DR, 3 - Hearing Protector Testing Room, 4 - the IEC standardized listening room in building 354, 5 - Lecture Room 019 in Building 352, 6 - The library in Building 352, 7 - Meeting Room 112 in Building 352.

Music Ratings					
SQ	Score	Boxiness	Score	Boominess	Score
1. HP	3.25	1. Library	2.56	1. HP	3.42
2. Talk	3.97	Meeting	2.61	(2.) Talk	3.53
3. Lecture	4.52	IEC	2.62	3. Control	3.67
4. Control	4.88	Talk	2.65	4. Meeting	3.94
5. IEC	5.63	2. Control	2.94	5. Lecture	4.47
Library	5.66	3. Lecture	3.38	(6.) Library	4.61
Meeting	5.69	4. HP	4.18	7. IEC	4.75
Speech Ratings					
SQ	Score	Boxiness	Score	Boominess	Score
1. HP	3.61	1. Talk	2.18	1. Talk	3.81
2. Lecture	4.44	2. Library	2.62	(2.) Control	3.92
3. Meeting	5.00	3. IEC	3.00	(3.) Meeting	4.06
4. Control	5.53	4. Meeting	3.24	(4.) HP	4.14
IEC	5.69	Control	3.29	(5.) IEC	4.22
5. Library	6.58	5. Lecture	3.50	(6.) Library	4.36
Talk	6.67	6. HP	4.09	7. Lecture	4.53
Total Ratings					
SQ	Score	Boxiness	Score	Boominess	Score
1. HP	3.43	1. Talk	2.41	1. Talk	3.67
2. Lecture	4.49	2. Library	2.59	2. HP	3.78
3. Control	5.21	3. IEC	2.81	Control	3.79
Talk	5.32	Meeting	2.93	3. Meeting	4.00
Meeting	5.35	4. Control	3.12	4. IEC	4.49
4. IEC	5.67	5. Lecture	3.44	Library	4.49
5. Library	6.13	6. HP	4.13	Lecture	4.50

Table 6.4: Room mean overall sound quality (SQ), boxiness and boominess ranking, for the music program ("Herbert") and speech ("Male") with both test parts averaged. The smallest rank designates the worst SQ, the least boxy and the least boomy (or the thinnest) rooms. Mean ratings that are clustered and are not significantly different share the same ranks. Ranks in parentheses signify that the mean is not confidently different than its higher and lower ranked rooms. Room abbreviations: HP - Hearing Protector Testing Room, Talk - Talk Studio 8 in DR, Control - DR Studio 3 Control Room, Lecture - Lecture Room 019 in Building 352, Meeting - Meeting Room 112 in Building 352, Library - in Building 352, IEC - the standardized listening room in building 354.

the measurements, as the large-room definition showed very poor correlation. To repeat the BR definition:

$$BR = \frac{T_{60}(125Hz) + T_{60}(250Hz)}{T_{60}(500Hz) + T_{60}(1000Hz)} \tag{6.9}$$

where the octave band T_{60}'s in the numerator are sometimes replaced with lower bands of 63 and $125Hz$. Only the latter is shown below in table 6.5, 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 slightly lower correlation with boominess. However, these figures are hardly satisfactory.

The T_{30} is the mean of the third-octave band data measured between $200 - 4000Hz$ (see for example [6]). The EDT is calculated in the same manner analogously. Another quantity that is examined is the recommended mean RT or T_m, as appears in the ITU and EBU standards for listening rooms [6] and [7]. Here the ratio between the actual RT and the recommendation is examined:

$$T_{30}/T_m = \frac{T_{30}}{0.25\left(\frac{V}{V_0}\right)^{1/3}} \propto \frac{T_{30}}{V^{1/3}} \tag{6.10}$$

Where V_0 is a reference volume of $100m^3$. An improvement is shown using a few specially optimized quantities. The Small room Bass Ratio was defined using the third-octave T_{30} values:

$$SBR = ln\frac{T_{30}(63Hz) + T_{30}(80Hz)}{T_{30}(250Hz) + T_{30}(315Hz)} \tag{6.11}$$

And the Small room EDT Bass Ratio, still using third-octave values:

$$SEBR = ln\frac{EDT(80Hz) + EDT(100Hz)}{EDT(250Hz) + EDT(315Hz)} \tag{6.12}$$

Where both quantities show better correlation if their logarithm is taken. The last new quantity introduced here is the Low-High Ratio (LHR), which was optimized especially to give a higher correlation with the boominess ratings:

$$LHR = ln\frac{T_{30}(50Hz) + T_{30}(63Hz)}{T_{30}(3150Hz) + T_{30}(4000Hz)} \tag{6.13}$$

A general remark has to be made 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 inspected values. Bearing all that in mind, we proceed to examine the strong correlations, validity and possible implications.

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

Some things can be generalized more safely. Boxiness can generally be considered an unwanted characteristic picked by listeners. The term boxiness was used in the first place as an indirect measure of coloration. However, the high correlations shown between boxiness and the RT and EDT, especially apparent in the speech ratings, casts a doubt over the connection 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. In large halls the similar bass ratio is associated with warmth and brilliance of the sound. Do these qualities have any association with the suggested SBR and SEBR? It cannot be inferred from the available data. The poor correlation of bass ratio with any rating might imply some different meaning.

It is likely that the embedded error in the SEBR is rather high, as it uses a 10dB slope to estimate the EDT at low frequencies. Nevertheless, its consistency between the three means and the similar trends of

Music Rating Correlation Matrix										
	SQ	Boxiness	Boominess	BR	SBR	SEBR	T_{30}	EDT	T_{30}/T_m	LHR
SQ	1.0000	-0.7873	0.7152	0.7399	**0.8359**	0.6204	-0.3951	-0.4504	-0.5557	0.6106
Boxiness	-0.7873	1.0000	-0.4315	-0.5117	**-0.8872**	**-0.8712**	0.8084	0.8484	0.8537	-0.5600
Boominess	0.7152	-0.4315	1.0000	0.7129	0.7338	0.5813	-0.1263	-0.1368	-0.3589	**0.8797**
BR	0.7399	-0.5117	0.7129	1.0000	0.7547	0.5792	-0.0821	-0.1310	-0.2738	0.8262
SBR	0.8359	-0.8872	0.7338	0.7547	1.0000	0.9088	-0.5994	-0.6520	-0.7741	0.8752
SEBR	0.6204	-0.8712	0.5813	0.5792	0.9088	1.0000	-0.8040	-0.8141	-0.8863	0.7064
T_{30}	-0.3951	0.8084	-0.1263	-0.0821	-0.5994	-0.8040	1.0000	0.9908	0.9444	-0.2400
EDT	-0.4504	0.8484	-0.1368	-0.1310	-0.6520	-0.8141	0.9908	1.0000	0.9627	-0.2977
T_{30}/T_m	-0.5557	0.8537	-0.3589	-0.2738	-0.7741	-0.8863	0.9444	0.9627	1.0000	-0.5184
LHR	0.7329	-0.5600	0.8797	0.8262	0.8752	0.7064	-0.2400	-0.2977	-0.5184	1.0000

Speech Rating Correlation Matrix										
	SQ	Boxiness	Boominess	BR	SBR	SEBR	T_{30}	EDT	T_{30}/T_m	LHR
SQ	1.0000	**-0.9611**	-0.3005	0.3383	0.7012	**0.8996**	**-0.9139**	**-0.8956**	-0.8511	0.3417
Boxiness	-0.9611	1.0000	0.3156	-0.3028	-0.7013	**-0.8916**	**0.8625**	**0.8594**	0.8146	-0.3513
Boominess	-0.3005	0.3156	1.0000	0.5610	0.2511	0.0854	0.4175	0.4090	0.1572	**0.6703**
BR	0.3383	-0.3028	0.5610	1.0000	0.7547	0.5792	-0.0821	-0.1310	-0.2738	0.8262
SBR	0.7012	-0.7013	0.2511	0.7547	1.0000	0.9088	-0.5994	-0.6520	-0.7741	0.8752
SEBR	0.8996	-0.8916	0.0854	0.5792	0.9088	1.0000	-0.8040	-0.8141	-0.8863	0.7064
T_{30}	-0.9139	0.8625	0.4175	-0.0821	-0.5994	-0.8040	1.0000	0.9908	0.9444	-0.2400
EDT	-0.8956	0.8594	0.4090	-0.1310	-0.6520	-0.8141	0.9908	1.0000	0.9627	-0.2977
T_{30}/T_m	-0.8511	0.8146	0.1572	-0.2738	-0.7741	-0.8863	0.9444	0.9627	1.0000	-0.5184
LHR	0.3417	-0.3513	0.6703	0.8262	0.8752	0.7064	-0.2400	-0.2977	-0.5184	1.0000

Total Rating Correlation Matrix										
	SQ	Boxiness	Boominess	BR	SBR	SEBR	T_{30}	EDT	T_{30}/T_m	LHR
SQ	1.0000	**-0.9188**	0.3779	0.6122	**0.8923**	**0.9009**	-0.7882	-0.8066	-0.8354	0.6106
Boxiness	-0.9188	1.0000	-0.1489	-0.4243	-0.8293	**-0.9221**	0.8743	**0.8933**	0.8722	-0.4750
Boominess	0.3779	-0.1489	1.0000	0.7051	0.6177	0.4517	0.0461	0.0357	-0.2094	**0.8627**
BR	0.6122	-0.4243	0.7051	1.0000	0.7547	0.5792	-0.0821	-0.1310	-0.2738	0.8262
SBR	0.8923	-0.8293	0.6177	0.7547	1.0000	0.9088	-0.5994	-0.6520	-0.7741	0.8752
SEBR	0.9009	-0.9221	0.4517	0.5792	0.9088	1.0000	-0.8040	-0.8141	-0.8863	0.7064
T_{30}	-0.7882	0.8743	0.0461	-0.0821	-0.5994	-0.8040	1.0000	0.9908	0.9444	-0.2400
EDT	-0.8066	0.8933	0.0357	-0.1310	-0.6520	-0.8141	0.9908	1.0000	0.9627	-0.2977
T_{30}/T_m	-0.8354	0.8722	-0.2094	-0.2738	-0.7741	-0.8863	0.9444	0.9627	1.0000	-0.5184
LHR	0.6106	-0.4750	0.8627	0.8262	0.8752	0.7064	-0.2400	-0.2977	-0.5184	1.0000

Table 6.5: Correlation matrices between all rating data and acoustical parameters in Experiment A for music, speech and both averaged. See text for definitions. The highest correlation between the dependent and independent variables is shown in boldface.

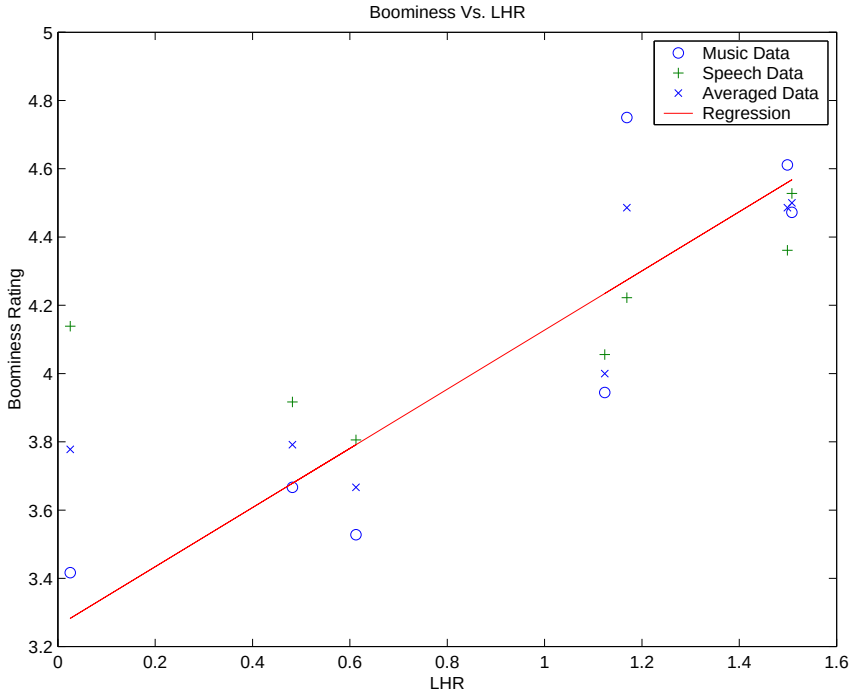


Figure 6.9: Mean data for the boominess vs. LHR. The regression line is shown for the music data. The regression yields $R^2 = 0.774$.

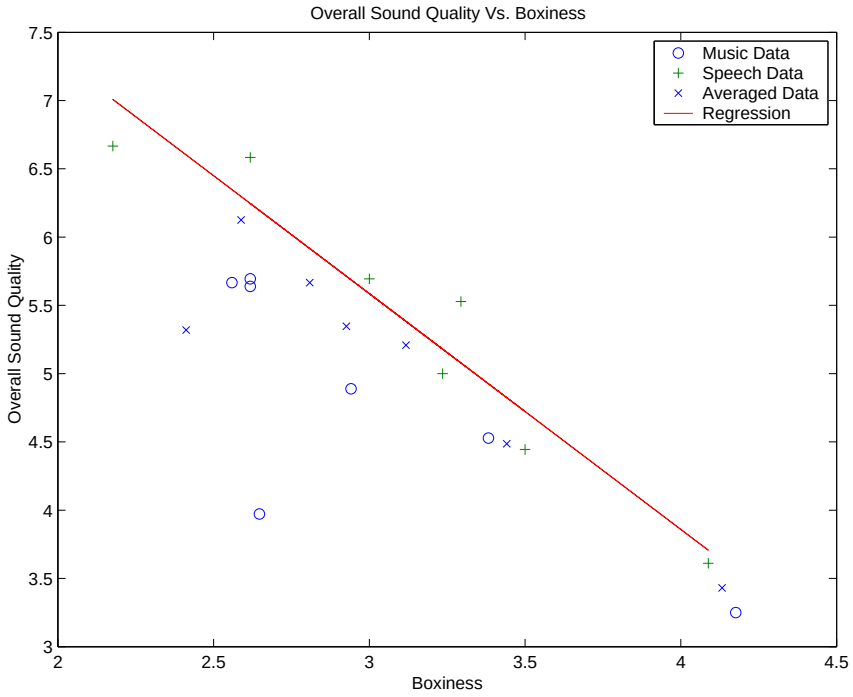


Figure 6.10: Mean data for the sound quality vs. boxiness. The regression line is shown for the speech data. The regression yields $R^2 = 0.936$.

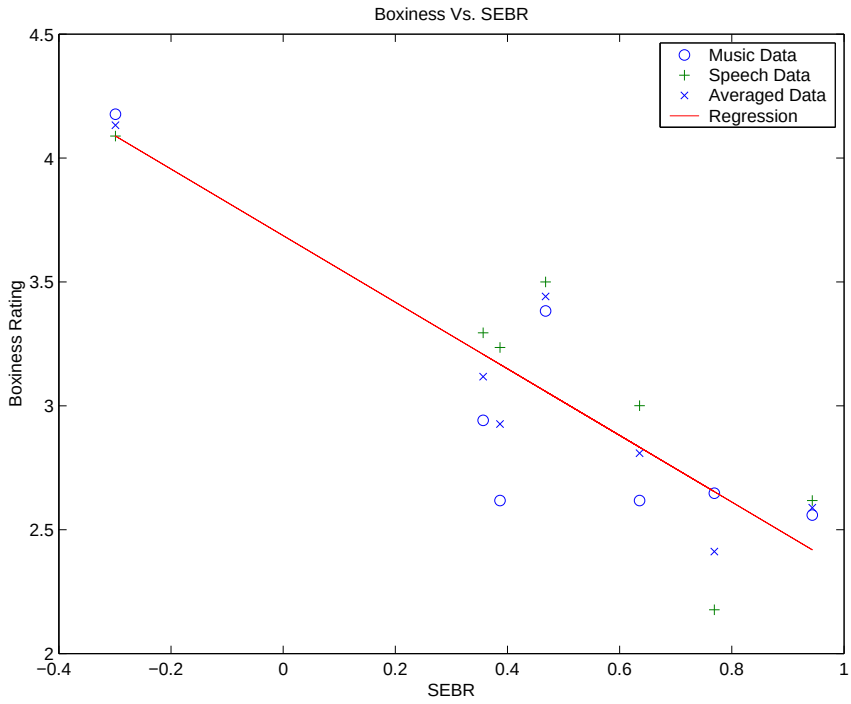


Figure 6.11: Mean data for the boxiness vs. SEBR. The regression line is shown for the averaged data. The regression yields $R^2 = 0.85$.

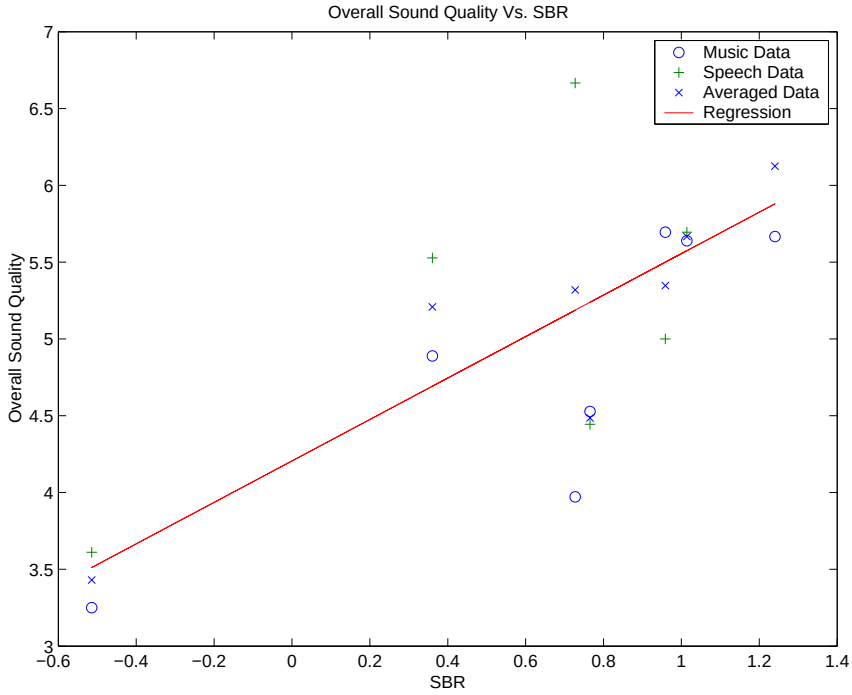


Figure 6.12: Mean data for the sound quality vs. SBR. The regression line is shown for the averaged data. The regression yields $R^2 = 0.796$.

SBR, which utilizes the more stable T_{30}, give some confidence. The normal bass ratio, which shows poor correlations, is computed using one-octave band values and is more precise.

The relation between T_{30} and SQ is shown in figure 6.13 Speech ratings show preference to the lowest RT's available (talk studio) and the library is the only room, which is preferred despite its higher mean T_{30}. The talk studio has already a very dead mean T_{30} of $0.12s$. Does the same trend continue outside the range - would the highest SQ be achieved in an anechoic room? One may speculate that the oppressive, dead nature of the anechoic room and recording would not be preferable. 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 two things. First, its relatively high volume combined with the rather dead acoustics for midrange and treble provides an unobtrusive environment. Second, it is a room, which was not acoustically designed, definitely not for the purpose used here, that can be perceived 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.

T_{30} in the music ratings shows a different trend. Although no specific function is fitted to the data (it is interpolated only for clarification), it is clearly seen in figure 6.13 that the SQ peaks at the narrow T_{30} range between 0.3 and 0.5s, quickly drops above and more slowly below that range.

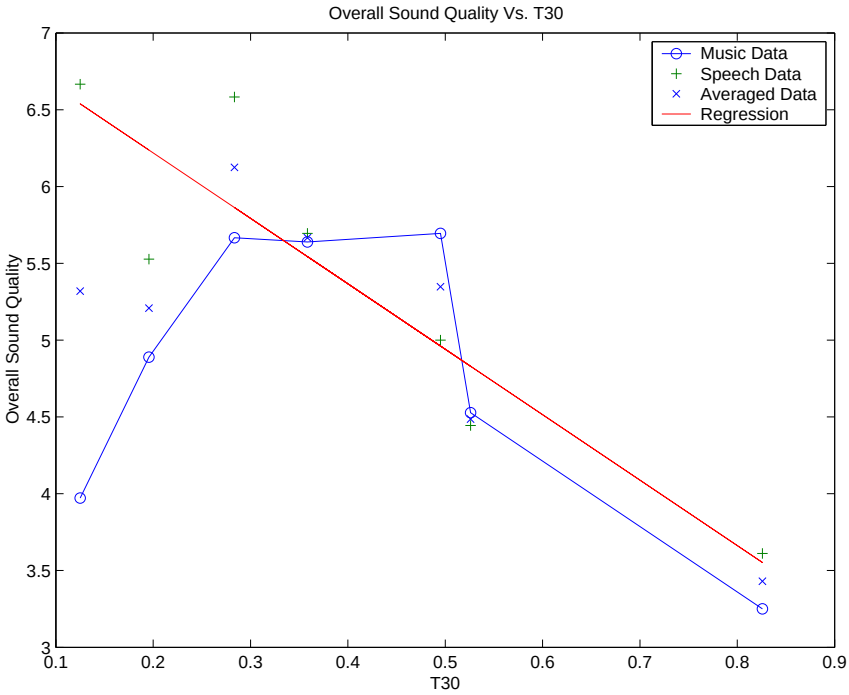


Figure 6.13: Mean data for the sound quality vs. T_{30}. The regression line is shown for the speech data. The music data is interpolated linearly. The regression yields $R^2 = 0.835$.

Boominess could not be well-correlated to anything but the newly introduced low-high ratio (LHR). Only one room, the hearing protector testing room, shifted the data from a monotonous 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 a hump in the midrange. It is possible that this in speech ratings it presented some masking, confusing or missing element to subjects, which made rating more random, averaging at the scale center of 4. However, the entire boominess rating showed very small variance and was very centered compared to the other two ratings in addition to lower reliability.

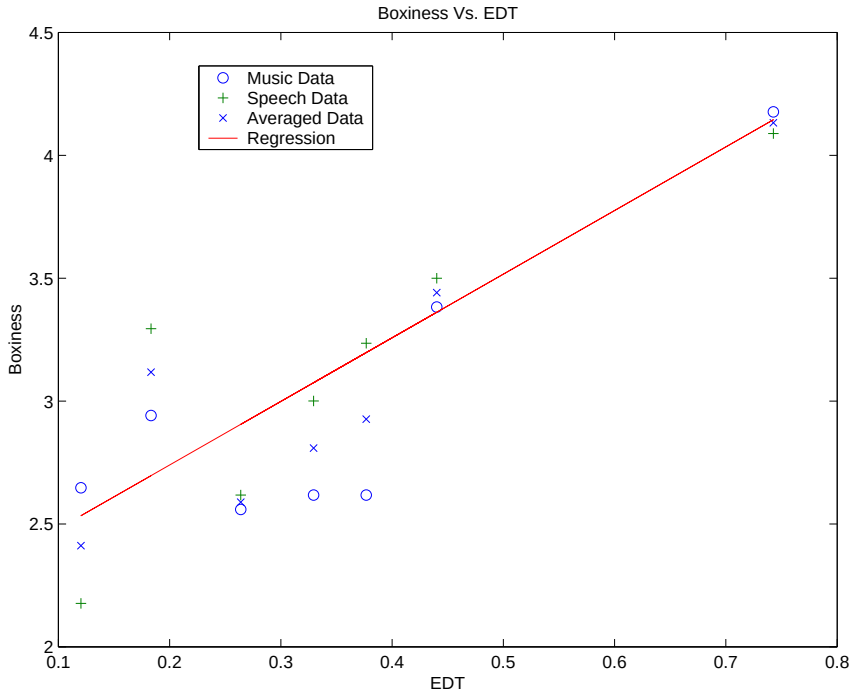


Figure 6.14: Mean data for the boxiness vs. EDT. The regression line is shown for the speech data. The regression yields $R^2 = 0.936$.

Hence its accuracy is probably not high. Accepting the LHR as a meaningful quantification of boominess, one may relate it to some tonal balance measure - low bass vs. treble, which can be perceived through the RT. Seeing that there in no clear functional dependence between the boominess and the SQ of the rooms (inspected for non-linear dependence as well, separately), this measure is of limited use. The lack of any correlation between SQ and boominess can be interpreted in a few ways: subject difficulty in understanding the boominess concept, wide tolerance for a 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.

6.3.2 Multiple Regressions

The last step that was taken was various multiple linear regressions between the SQ, boxiness and boominess and more than one acoustical variable at once. The results for speech and music are, yet 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:

$$SQ = b_1 \cdot SEBR + b_2 \cdot LHR + b_3 = 3.66 \cdot SEBR - 1.16 \cdot LHR + 4.73 \tag{6.14}$$

A graphical representation of the fit with the measured data is shown in figure 6.15. Interpolation between the points defines a plane. As was said before regarding the T_{30}, it is 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 the RT, the r^2 might even seem exaggerated. Using the same two independent variables, boxiness can be predicted as well, with $r^2 = 0.9498$.

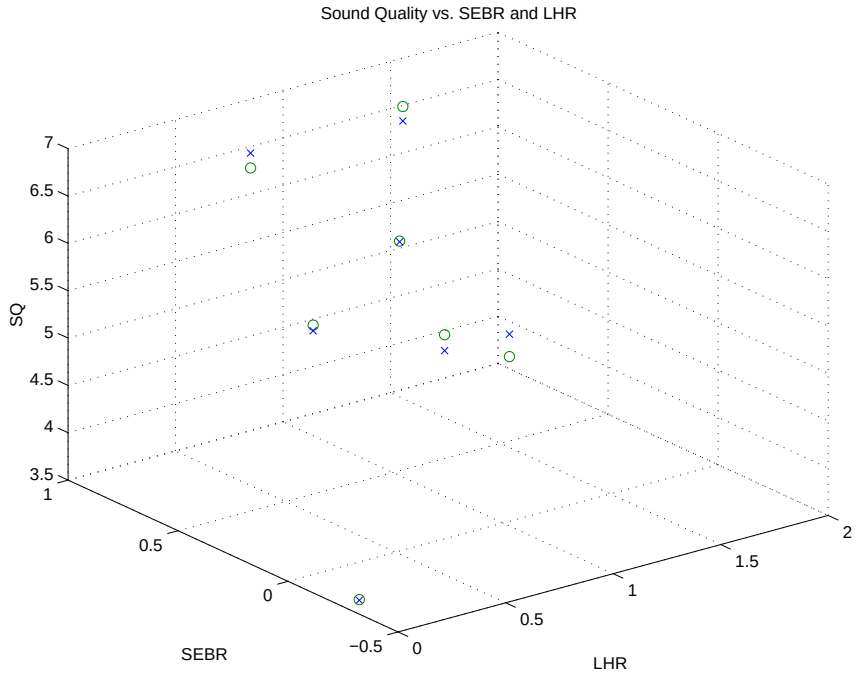


Figure 6.15: Measured and predicted overall sound quality ratings as a function of SEBR and LHR for speech. The multiple regression yields an $R^2 = 0.9815$.

Music program shows a more complicated trend. No less than 5 independent variables are needed to linearly model the SQ ratings for the music, a very dubious fit for only 7 points, which is therefore rejected. Multiple regression between SQ rating and boxiness and boominess, shows the best bit for speech, with $r^2 = 0.9236$ and considerably worse one for music with $r^2 = 0.7931$. Therefore, either there are hidden parameters, which were not gathered in this experiment, or that a linear model assumption does not hold. Most likely, it is both combined. The T_{30} example above supports a more general non-linear dependence with SQ.

6.3.3 Quality Summary of the Rooms

A summary of all the parameters mentioned above for all seven rooms is given in below in table 6.6.

It is not intended in this project to criticize or modify the particular rooms used, but some observations are suggested at any rate, in light of all of the above. Indirectly they serve to enhance the validity of the experiment.

All rooms can be tested for the standard listening room guidelines [7] and [6], which specify the mean RT, T_m (see eq. 6.10), and its tolerance above $200Hz$ and the RT slope at frequencies below that (see figure 6.16), although only the IEC listening room is used as a listening room per se:

- Only two rooms have their mean RT in the nominal range between $0.2s$ and $0.4s$ [7]: the IEC listening room and the library.
- The control room is nearly within the midrange 0.05s tolerance range of the volume dependent T_m, being a bit more dry ([7] and [6]). The library satisfies the recommended T_m, being only $0.02s$ apart The IEC listening room is not within the T_m tolerance limit.

- All rooms are still within the mean RT+0.3s bass slope tolerance at $80 - 125Hz$, but only the talk studio and the control room achieve that at $63Hz$.

- Only the IEC listening and the hearing protector testing rooms adhere to the low noise criterion of under 15 NR.

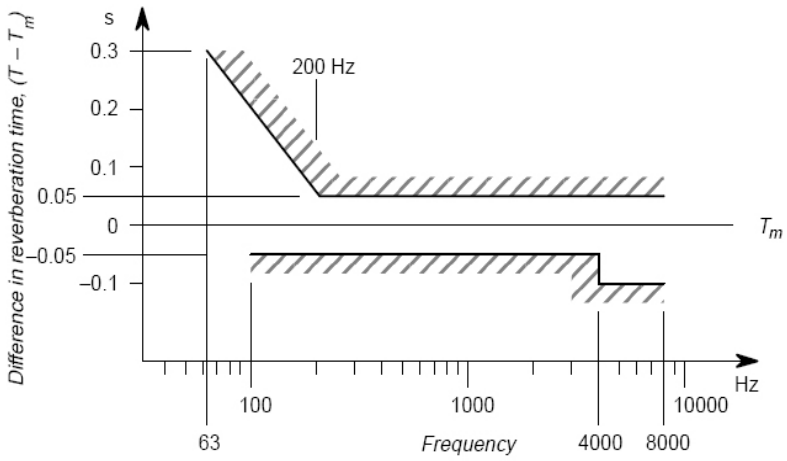


Figure 6.16: Tolerance curve of the RT in a listening room according to the EBU standard [7]. See the definition to T_m in eq. 6.10.

Information about the so-called "operational room response curve" in the rooms was not calculated and is not available. [2]

DR Talk Studio 8

The specially designed talk studio can be considered well-designed. It serves its purpose well, being the preferable room for speech reproduction compared to rooms with other functionalities. However, the very short reverberation time seems to be inadequate for a music listening for the average listener. It consistently sounded non-boxy, a fact that we like to relate to its irregular shape, which tackles coloration, yet the low RT is highly correlated with low boxiness and this relation is uncertain. Finally the room was also consistently non-boomy and that is predicted by use of the LHR parameter.

Library

The library scored well in all tests and can be thus considered a multi-purpose room, in a sense similar a to multi-purpose halls and auditoria. It enjoys both low mean RT, but its bass RT ensures a non-dead environment. The room size is an advantage for many purposes, but the particular use of space in the room is unpractical, as it is almost too crowded for anything but books.

[2]The standards define the operational room response curve as the deviation from a $\pm 3dB$ flat SPL response of the third-octave filtered pink noise played by the monitor speaker at the listening position in the room. This tolerance is valid to the $250 - 2000Hz$ range ($50 - 2000Hz$ in EBU) , but allows gradually sharper dips below (only ITU) and above it.

Control Room for DR Studio 3

The average performance of the control room is somewhat disappointing, considering its dedicated design and purpose. It seems that listeners tend to prefer a more reverberant bass (higher SBR), as in other rooms, maintaining an RT that is not too long.

It is interesting to note that a producer/sound engineer, who works regularly in this room expressed his dissatisfaction from the sound in the room, as it lacks in bass, no matter what loudspeakers are used.

IEC Listening Room

The listening room performed well in all tests. The room does not exactly comply with two other standards than the one according to which it was designed, yet it serves its purpose well as such, being adequate for both speech and music.

Meeting Room

This is another room, which can be viewed as a more natural and common environment. It offered little damping, as all walls are bare and there are hardly any furniture in the room. Its high ratings with the music program are rather surprising. A flutter echo was easily heard in the empty room upon hand clapping, which suggests high coloration. And yet it seemed secondary to the average rater, perhaps because of optimal room size or satisfactory bass sound and mean RT. However, the rather low speech rating suggest that for its more common use, for meetings and talks, it is not optimal. Of course, the demands from a meeting room should be nowhere as strict as a talk studio or a listening room.

Lecture Room

The poor performance of the lecture room was expected. A very big room, which is hardly treated, and is completely bare of absorbing furniture. The measurements were done with no people in the room - an uncommon situation. And yet there is so much that human absorption can offer. The room is regularly used for lectures, frequently supported by audio demonstrations. It can be safely said that the room is not optimized for these purposes. Although the mean RT in the room is comparable with the one in the meeting room, its bass RT runs too "wild" and the strong coloration (a strong flutter echo was noticed here as well) may no longer be dismissed.

Hearing Protector Testing Room

The worst rated room was fortunately so. It was not intended for any recreational or accurate sound reproduction, but to be as reverberant as possible, for very different purposes than were tested here. Its low scoring serves as an anchor to give some reassurance to other ratings, for a higher rating would have been alarming at least.

Parameter	Control Room	Talk Studio	HP Room	IEC Room	Lecture Room	Library	Meeting Room
V (m^3)	105	82	27	97	186	186	85
T_{30}(s)	0.1956	0.1250	0.8259	0.3584	0.5261	0.2835	0.4951
T_{20}(s)	0.2001	0.1254	0.8259	0.3646	0.5344	0.2826	0.4941
EDT (s)	0.1834	0.1204	0.7426	0.3294	0.4401	0.2639	0.3767
f_s (Hz)	87	79	350	122	107	78	153
$BR_{(63,125Hz)}$	1.173	1.360	1.175	2.042	2.118	3.449	2.658
$BR_{(125,250Hz)}$	1.012	0.945	1.434	1.047	1.119	1.407	1.273
SBR	0.3607	0.7271	-0.5142	1.0141	0.7654	1.2405	0.9590
SEBR	0.3567	0.7688	-0.2989	0.6354	0.4680	0.9435	0.3866
T_m (Hz)	0.2541	0.2340	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	0.4820	0.6127	0.0254	1.1690	1.5082	1.4990	1.1237

Table 6.6: Room acoustical data: V - approximate volume, mean T_{30}, T_{20} and EDT between $200-4000Hz$, f_s - approximate Schröder's frequency, two alternative bass ratios using different octave bands in the numerators. See text for the other definitions, section 6.3.1.

Chapter 7

Test Results and Analysis - Experiments B and C

Due to the procedural similarities of Experiments B and C, some of their analyses are brought together, to avoid repetition.

7.1 Task Fulfillment

Experiment B

Experiment B was the opening listening test. While the task definition may have seemed odd to a number of examinees in the start, through the first simple example it was immediately made clear, as the room acoustics difference in the two pairs was so obvious. The second example was more difficult and not everybody answered it correctly. Most subjects showed improvement in the retest. Only two subjects in the first part showed 6 correct answers - which may still be insignificant (above 5%). On the other hand, some subjects performed remarkably well and one was able to match all pairs correctly in both parts.

A few test subjects noted that speech samples were much easier to discern than music. Also, they noted that the musical program used ("Mingus"), in large ensemble jazzy style, was the most difficult to match.

Finally, it was noted that quite a few subjects answered correctly in the first part a certain question, but incorrectly in the other. That may indicate an especially confusing pair, but also the successful employment of randomization in the test. As no subject (including the author) was able to remember a specific set, but had to perform the task anew.

Experiment C

Explanation of the task in Experiment C posed no problem, yet some subjects were taken aback already by the difficulty of the "easy" example. Some reverted to some kind of (intelligent) guessing in that stage.

In general, the relative scores were considerably lower than before and were closer to chance probability, as is shown below. Test subjects said how it is extremely difficult to compare speech and music signals. Also some signals are much more difficult to decipher such as "Mingus" and 'Machaut", the latter was presented only in the examples. With all that taken together it is understandable that many subjects showed poor repeatability between the two parts of the test.

Two things should be noted about individual scores that were noticed immediately. First, most subjects who performed especially well in Experiment B, scored poorly in Experiment C. Second, some individuals scored remarkably well and consistently. Above 6 out of 8 is the significance margin for the individual

test and it was reached twice, both times by subjects who performed poorly on the first part. One more particularly confusing case was of a subject who scored 7 out of 8 in the first part, but inconsistently dropped to 2 out of 7 in the second. He was asked to repeat the test and then he scored 5 out of 8 followed by 2 out of 8 again. While not impossible probabilistically, these results are rather confusing.

All three examinees were interviewed later and their remarks will be discussed in a later section. As a preliminary pointer, it seemed that since that the two tests correlated so poorly amongst the same subjects, there must be at least one different factor in each test, which does not help to succeed in the other.

7.2 Test Analysis

The following things are of interest in the first part of the analysis: obtaining a score for subjects, corrected for chance probability; examination of a possible learning effect; setting probabilities for correctly answering particular items and inferring a relative corresponding difficulty level. Also the reliability of the tests will be examined.

Later stages include comparison of particular questions, which exist in both experiments with variations. Furthermore, some prediction might be suggested for the difficulty levels of questions, based on the room acoustics and other factors.

7.2.1 Test Scores

The final score output from each subject showed 1 for a correctly answered question and 0 for a wrong answer. Summing up the total correct answers in every test gives a total score per subject. Being derived from a multiple-choice test, with no "I don't know" alternative, the total score has to be corrected for chance. In Experiment B the guessing average is 4 out of 12 and in Experiment C, $2\frac{2}{3}$ out of 8. Taking that into account the guessing average is defined as 0 score. There are 4 possibilities per question per subject in each test, looking at its two parts: wrong answers in both parts, correct only in the first part, correct only in the second part and correct in both parts. Hence there is one more correction, which may arise if a subject shows what can be viewed as a "negative learning" effect, in case when one answers a question correctly only in the first part. This display of poor consistency with some subjects, especially in Experiment C, may indicate either possible guesswork or at least no confident knowing of the correct answer. In the other case, where a subject answers correctly only in the second part, there is still a chance for guesswork, but also higher probability of some "positive learning". These considerations were taken into the corrected score formula:

$$S' = 2 * (S_1 + S_2) - (W_1 + W_2) - C_1 \tag{7.1}$$

Where S' is the corrected score, S_i is the total correct in part i, W_i is the total wrongs in part i and C_1 is the total items that were answered correctly only in part 1. Both corrections allow negative scores as well, but the whole scale can be then transposed to start from 0. It will not be done here though, to maintain a mutual reference point (0) for both experiments. The summary of all tests (author's included) is illustrated in figures 7.1 and 7.2. The original and corrected scores are given in table 7.1.

It should be noted that when looking at the uncorrected individual test scores, a significance level can be found to answer the question: What is the chance that the subject was guessing through the entire test? Binomial distribution is used here and for the values of $p = \frac{1}{3}$, $q = \frac{2}{3}$ and $n = 12$ or $n = 8$, it is easily obtained that scores of 7/12 and 6/8 are there is less than 5% probability to get through chance, whereas 9/12 and 7/8 are less than 1%. The 5/8 score is has less than 7%. The binomial coefficient is calculated according to:

$$s_{k/n} = \mathrm{C}_n^k p^k q^{n-k} \tag{7.2}$$

where $s_{k/n}$ is the probability for getting k/n answers right by chance.

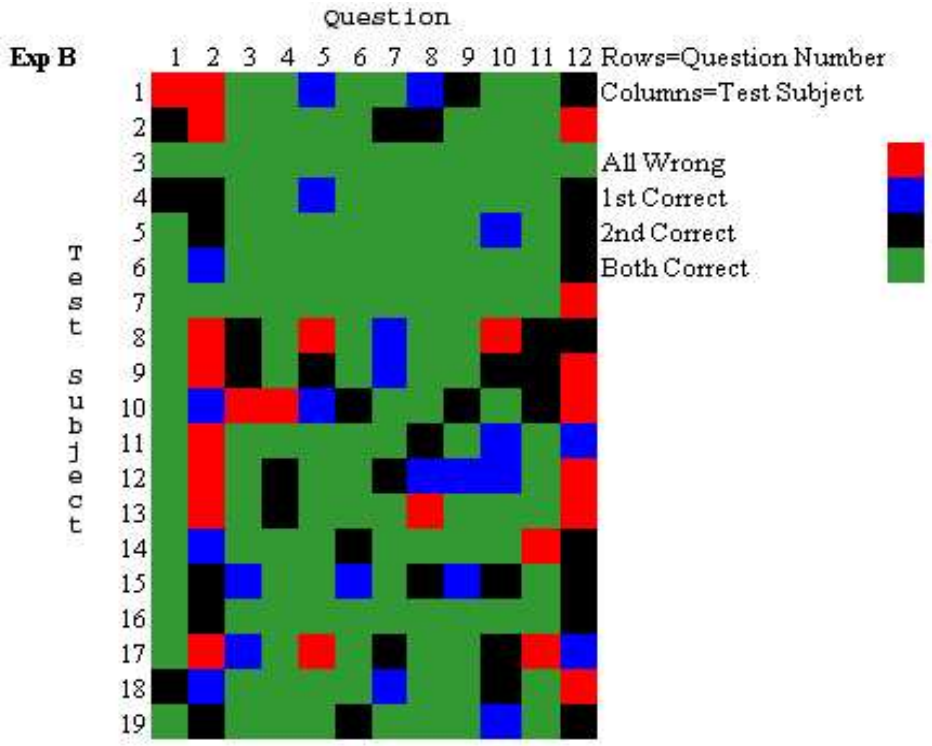


Figure 7.1: Individual Scores in Experiment B. Rows stand for subjects and columns stnad for test items. The color of their intersection shows the type of individual success per item per subject, according to the color legend. Note that the subject numbers do not correspond to figure 7.2. See table 7.1 for the matched scores.

Rater	Exp B - S1	Exp B - S2	Corrected B	Exp C - S1	Exp C - S2	Corrected C
1	8	8	22	3	3	1
2	7	10	27	3	2	-4
3	12	12	48	3	4	5
4	9	11	35	3	4	5
5	10	11	38	2	3	-2
6	11	11	41	3	4	4
7	11	11	42	1	3	-4
8	6	8	17	4	3	4
9	7	9	23	4	3	3
10	6	7	13	5	6	17
11	10	9	31	7	2	6
12	8	7	18	3	6	11
13	8	9	27	2	1	-9
14	9	10	32	2	4	2
15	8	9	24	2	4	1
16	10	12	42	4	4	7
17	7	8	16	2	2	-6
18	9	9	28	3	2	-2
	9	11	31	3	5	8

Table 7.1: Individual Correct-Answer and Chance-Corrected Scores in Experiments B and C

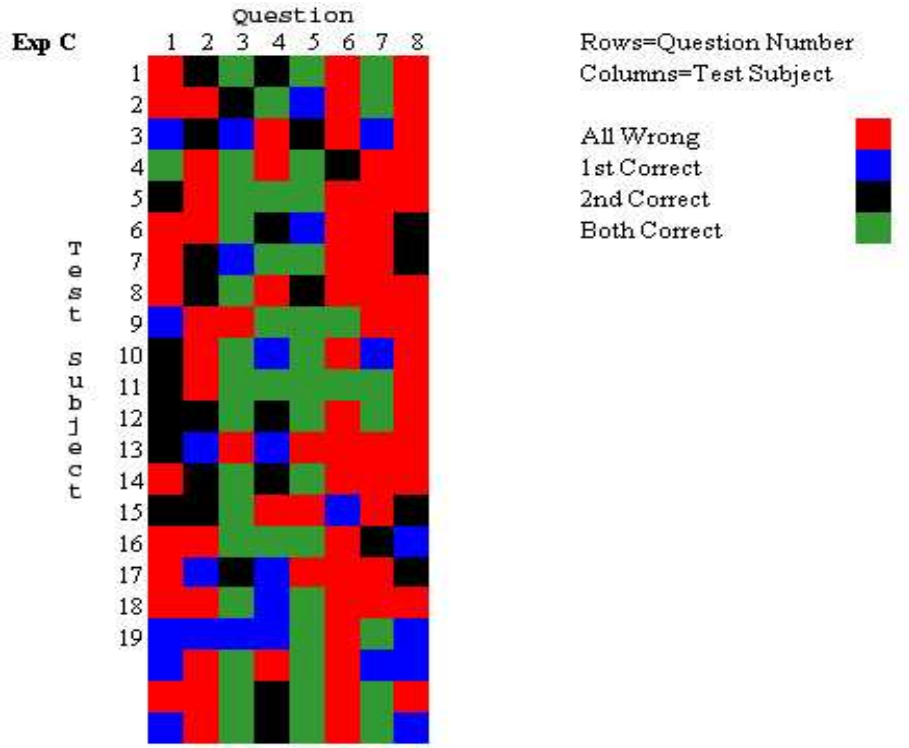


Figure 7.2: Individual Scores in Experiment C. Rows stand for subjects and columns stnad for test items. The color of their intersection shows the type of individual success per item per subject, according to the color legend. Note that the subject numbers do not correspond to figure 7.1. See table 7.1 for the matched scores.

7.2.2 Learning Effect

Learning effect was mentioned before in relation to the corrections made. It is referred to, in these experiments, to an improvement in subject capability in the repetition round, due to training and experience gained in the first round. If such is the case we can expect higher subject scores in the second part, as indeed is the case with many subjects. However, as a seemingly negative learning effect occasionally appears, where specific items are not consistently answered correctly, that learning has to be examined. A paired t-test was performed on the two sets of data from each part in every experiment. The test examines what significance does the hypothesis have, which says that the mean score in part 2 is higher than in part 1. T-tests show a learning effect significance of 9% in Experiment B and 6.5% in Experiment C. That is, it is likely to take place, but it cannot be stated confidently with the usual 5% significance level or lower, as not all subjects had shown learning and chance might have played a role in some individual total score improvements.

7.2.3 Item Difficulty Levels

Delving a little deeper in the test, we are interested in establishing item-specific data, namely what is the success rate per test item. Binomial distribution can be exercised once again to examine the mean obtained per test item, against its expected guessing chance. This item difficulty is illustrated in figures 7.1 and 7.2 by looking at the columns. Columns with dominant green color show high scores and consist all but 2 questions in Experiment B. The situation in markedly different in Experiment C, where only two columns are green-dominated and most of the others show much more red - all wrong - scores. Item probability becomes interesting if it can tell us something about how difficult it is and in turn, reasons for relative ease or difficulty could be sought. Although the probability of items is rather clear as scoring percentage and as an index of difficulty (actually it directly describes easiness and not not difficulty), it has to be corrected for chance as well. Also, it has to be transformed to a linear quantity, as the underlying assumption is that the difficulty of an item is related to probability through cumulative normal distribution or normal-ogive function (see figure 7.3). This assumption then makes use of the z value of the ogive, in order to transform p into a linear difficulty scale [41]. Let us review the necessary formulae. The standard deviation σ_i, of an item i is needed for its z value calculation:

$$\sigma_i = \sqrt{p_i q_i} \tag{7.3}$$

Where p_i is the proportion of passing the item and $q_i = 1 - p_i$ is the proportion not passing the item. It can be seen that σ_i is maximal when $p = q = 0.5$. Therefore, that is the point on the curve which has the most variance and is chosen as the reference difficulty 0.

Taking chance into account in the probability p_i, a correction formula by Guilford [41] is applied:

$${}_cp_i = \frac{kp_i - 1}{k - 1} = \frac{3p_i - 1}{2} \tag{7.4}$$

where ${}_cp$ is the corrected probability and k is the number of alternatives in the test, in our case 3. At the chance probability the corrected value becomes 0. Below the chance probability the formula assumes negative values, so it is not applicable and ${}_cp_i$ is set to 0.

Finally the z_i value, or an item difficulty of a test item, is given by:

$$z_i = \frac{x}{\sigma_i} = \frac{{}_cp_i - 0.5}{\sigma_i} \tag{7.5}$$

where x is the probability deviation from 0.5, the mean of the ogive, which is defined as 0 difficulty.

It can be seen in the table that a few items showed less than 1/3 chance to be answered correctly. If that uncorrected probability is the true mean, then it may indicate a systematic bias in these test items. In other words, questions that have lower probability than chance, can be considered misleading,

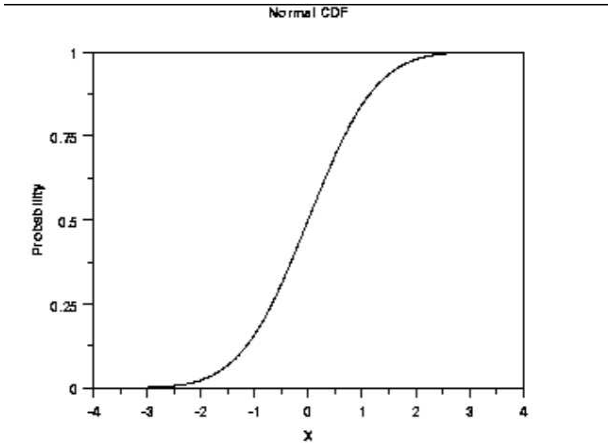


Figure 7.3: A cumulative normal distribution function

Item	p_1	${}_cp_1$	$\sigma 1$	$z1$	p_2	${}_cp_2$	$\sigma 2$	$z1$
B1	0.7778	0.6667	0.4157	0.4009	0.9444	0.9167	0.2291	1.8190
B2	0.3333	0	0.4714	-1.0607	0.3333	0	0.4714	-1.0607
B3	0.8889	0.8333	0.3143	1.0607	0.8333	0.7500	0.3727	0.6708
B4	0.8333	0.7500	0.3727	0.6708	0.9444	0.9167	0.2291	1.8190
B5	0.8333	0.7500	0.3727	0.6708	0.7222	0.5833	0.4479	0.1861
B6	0.8889	0.8333	0.3143	1.0607	0.9444	0.9167	0.2291	1.8190
B7	0.8333	0.7500	0.3727	0.6708	0.8333	0.7500	0.3727	0.6708
B8	0.7778	0.6667	0.4157	0.4009	0.8333	0.7500	0.3727	0.6708
B9	0.8889	0.8333	0.3143	1.0607	0.8889	0.8333	0.3143	1.0607
B10	0.7222	0.5833	0.4479	0.1861	0.7778	0.6667	0.4157	0.4009
B11	0.7222	0.5833	0.4479	0.1861	0.8889	0.8333	0.3143	1.0607
B12	0.1667	0	0.3727	-1.0607	0.5000	0.2500	0.5000	-0.5000
C1	0.2222	0	0.4157	-1.0607	0.3889	0.0833	0.4875	-0.8547
C2	0.1667	0	0.3727	-1.0607	0.3333	0	0.4714	-1.0607
C3	0.7778	0.6667	0.4157	0.4009	0.7222	0.5833	0.4479	0.1861
C4	0.6111	0.4167	0.4875	-0.1709	0.5000	0.2500	0.5000	-0.5000
C5	0.7222	0.5833	0.4479	0.1861	0.7222	0.5833	0.4479	0.1861
C6	0.1667	0	0.3727	-1.0607	0.1667	0	0.3727	-1.0607
C7	0.3333	0	0.4714	-1.0607	0.2778	0	0.4479	-1.0607
C8	0.1111	0	0.3143	-1.0607	0.2222	0	0.4157	-1.0607

Table 7.2: Summary of item probability p_1, corrected probability for chance ${}_cp$, standard deviation σ and difficulty z in parts 1 and 2 in Experiments B and C

where one or both wrong alternatives are more appealing to test subjects than the correct alternative. Unfortunately, the data saying which alternative did the subject choose in each question was not recorded and an exact examination cannot be carried out.

The corrected probabilities are not going to be used in later calculations, but they do give a further insight on the test items.

7.2.4 Test Reliability

The test-retest structure of the experiments allows a simple check of reliability, or more accurately, of stability of the two tests. It is telling of how dependent the subjects are on time (between tests) and how stable are the measurements. Generally, due to chance scores, multiple choice test are deemed less reliable than other testing methods. The more alternatives there are per item in the test, its reliability increases. Computation of reliability in this case would be simply the coefficient of correlation between the two test parts.

Experiment B shows $r_{tt} = 0.7868$ between the two test administrations. Going back to figure 7.1, it shows that many of the subjects are rather consistent for most of the test items, but taking into account the learning effect suggested before, it is obvious that the test reliability will drop as a result.

Experiment C, on the other hand, shows a much poorer reliability of $r_{tt} = 0.4103$, which is understandable due to considerably higher difficulty of all test items and resorting to what seemed as guessing for the most part. In fact, one subject, who scored 7/8 and 2/8 in the two test parts, single-handedly contributes to the poor test reliability. Leaving his scores effective, the reliability drops to $r_{tt} = 0.108$! Therefore his scores can be exchanged with the author's scores, whose acquaintance with the test material can be assumed insignificant, due to the high randomization of this test.

7.2.5 Discrimination Mechanisms

There are various acoustical factors that may contribute to subjects' capability to distinguish between rooms in both experiments. The most obvious one is the difference between RT characteristics of the two rooms. Another is the difference in the source-receiver coupling functions (see eq. 2.8), which is determined by the room shape. Dissimilarity between two samples, which were recorded in the same room, is higher if both terms in 2.8 are altered, i.e. both receiver and source were moved. It must be taken into consideration that there may be minute variations in the mean RT within a room - over recording positions, which gradually increase the lower the frequency gets.

The subjects who fared better in Experiment C, were asked what was the element which they listened to, which helped the most. They referred to the room size, volume and to its boundaries which can be sensed. One subject (no. 12 in table 7.1) said that the training and concepts gained by Experiment A, such as boxiness and boominess, made him focus on them and better discriminate the correct pair. However, other subjects, when asked, claimed to have done the same (room size) - apparently with little success. The best subject in Experiment C (no. 10 in table 7.1), who had the worst total score in Experiment B, was asked to redo one part of it, in order to see if there is an improvement and if so, to what extent. This time he scored 10/12 an improvement over 7/12, yet not a perfect score. The best subject in Experiment B, who had no mistakes, said that although he focused on the bass notes, in difficult situations he relied on the binaural effect and tried to visualize the room.

Taking all these into account together with musical program (speech/music/mixed) and perhaps learning effect as well, forms a rather complicated picture.

The data in table 7.2 represents the complete set of statistical dependent data available from the experiments. Relating it to any other independent parameter might help us to form a prediction model. Since the dependent variable is a probability, or namely, success or failure in the task of distinguishing between the two rooms in a particular test item, the prediction model must be logistic. Logistic regression model fits independent parameters to a hypothesized S-shaped function, whose dependent values are probabilities

and hence constrained between 0 and 1. Logistic regression is suitable for binomial distributions and does not make any assumptions on the set of data, such as normality of the samples.

Going back to table 7.2, it can be seen that the probabilities p_1 and p_2 are rather unevenly distributed. ${}_cp_1$ and ${}_cp_2$ are linear transformations of the original probabilities and behave very much the same. The main problem to be solved is what independent variable or variables can account best for the variance in probabilities of the different pairs.

Experiment B

At first, a means of comparing the two RT curves of every pair of rooms is necessary. Correlation to some degree between that and the task difficulty was predicted and should be tested. A few options were explored, using the one-third octave band room data of the EDT, T_{20} and T_{30} (see figures 4.21, 4.22 and 4.23):

$$X = \sum_i^N |RT_{1i} - RT_{2i}| \tag{7.6}$$

$$X = \sum_i^N |RT_{1i} - RT_{2i}|^2 \tag{7.7}$$

$$X = \sum_i^N \left|RT_{1i}^2 - RT_{2i}^2\right| \tag{7.8}$$

$$X = (SS_{BG}) = N(\overline{RT_1} - \overline{RT_T})^2 + (\overline{RT_2} - \overline{RT_T})^2 \tag{7.9}$$

In all equations X describes some RT difference between rooms 1 and 2, RT_i was substituted with either EDT, T_{20} or T_{30}. N is the total number of one-third octave bands measured. The last relation is obtained after one-way ANOVA between the two one-third octave vectors, which seemed as a reasonable method of comparing the two means. Although the prediction function must not be linear, the correlation coefficient was still used to obtain a rough estimate of the possible goodness of fit, as sets of data would become clustered when the independent variable is effective. It was found that the *log* of eq. (7.6) gives the highest correlation with the probabilities, using EDT or T_{30}, whereas the other equations showed much poorer correlations. Next, the correlation was tested for a narrower group of frequency bands. Do all bands count in the discrimination process? If so, are they weighted differently? All RT curves of the rooms show the biggest variances at the lowest frequencies, before flattening out at higher bands. Therefore, it is hardly surprising that this region is the dominant for the subject's choice. The correlations for various combination of frequency bands are summed up in table 7.3.

High correlation peaked at surprisingly low frequency range of T_{30} - between $50-63Hz$ - at $r = 0.8966$. Adding or subtracting terms (frequency bands) from the sum in eq. (7.6) either did not affect the correlation or worse, had it deteriorated to a lower value. The highest peak was found in the $50-160Hz$. Very high correlation of $r = 0.9014$ (for both sets of data combined) was obtained with only one term in the sum, $50Hz$. Other single bands proved much worse and all in all, the logistic fit seemed most convincing for the $50-160Hz$ combination, although high correlations were maintained for all combinations of 40, 50, 63 and 80 Hz bands. EDT showed a bit different numbers with correlation peaking also with the four-band combination, $80-160Hz$ ($r = 0.8684$), giving the best fit . Use of T_{20} showed similar behavior to T_{30} summing the differences in the bands $50-125Hz$ (r=0.8614).

Let us recapitulate the above by defining the the Bass T_{30} Difference (BTD), Bass T_{20} Difference (BTD2) and Bass EDT Difference (BET) as follows:

$$BTD = \ln(\sum_{f=50Hz}^{160Hz} |T_{30,Room1}(f) - T_{30,Room2}(f)|) \tag{7.10}$$

	T_{30}			EDT			T_{20}		
Frequency (Hz)	part 1	part 2	parts 1+2	part 1	part 2	parts 1+2	part 1	part 2	parts 1+2
40	0.7910	0.8215	0.8381	0.3778	0.5276	0.4624	0.5563	0.5766	0.5889
50	**0.8729**	**0.8554**	**0.9014**	0.7043	0.8646	0.8074	0.8468	0.8506	0.8839
63	0.6701	0.5958	0.6640	0.6572	0.7336	0.7197	0.5513	0.5006	0.5510
80	0.4759	0.4949	0.5045	0.6174	0.7237	0.6920	0.6692	0.6825	0.7032
100	0.4274	0.3953	0.4305	0.6680	0.4489	0.5952	0.5104	0.4988	0.5264
40-50	0.8505	0.8543	0.8878	0.6186	0.7505	0.7050	0.7806	0.7749	0.8106
50-63	**0.8573**	**0.8648**	**0.8966**	0.7054	0.8178	0.7865	0.7968	0.8084	0.8354
63-80	0.6743	0.6280	0.6812	0.7242	0.8082	0.7930	0.6582	0.6493	0.6816
80-100	0.5178	0.5225	0.5416	0.6747	0.6450	0.6892	0.6601	0.6635	0.6892
100-125	0.7485	0.6080	0.7152	0.6440	0.4235	0.5696	0.7835	0.6231	0.7425
125-160	0.8768	0.6408	0.8050	0.4945	0.3043	0.4277	0.6430	0.4699	0.5903
50-80	0.8486	0.8643	0.8913	0.7344	0.8452	0.8160	0.7932	0.8069	0.8326
50-100	0.8566	0.8649	0.8962	0.7764	0.8465	0.8411	0.8060	0.8139	0.8433
50-125	**0.8777**	**0.8607**	**0.9066**	**0.8210**	**0.8495**	**0.8684**	**0.8389**	**0.8105**	**0.8609**
50-160	**0.8923**	**0.8543**	**0.9121**	**0.8342**	**0.8420**	**0.8726**	0.8318	0.7984	0.8512
63-160	0.7896	0.6866	0.7753	0.8416	0.7791	0.8480	0.7439	0.6739	0.7429
80-160	0.8422	0.7139	0.8184	0.8558	0.7269	0.8323	0.8439	0.7395	0.8312
100-160	0.8895	0.7018	0.8405	0.7954	0.5690	0.7246	0.7779	0.6360	0.7452
100-200	0.8711	0.6792	0.8193	0.8343	0.6050	0.7638	0.7732	0.6243	0.7371

Table 7.3: Summary of item probability p_1, corrected probability for chance ${}_cp$, standard deviation σ and difficulty z in parts 1 and 2 in Experiments B and C

$$BTD2 = \ln(\sum_{f=50Hz}^{125Hz} |T_{20,Room1}(f) - T_{20,Room2}(f)|) \tag{7.11}$$

$$BED = \ln(\sum_{f=50Hz}^{160Hz} |EDT_{Room1}(f) - EDT_{Room2}(f)|) \tag{7.12}$$

The frequency ranges in these expression are more stable than the lower ones, which also show high correlations, as they are very close to each other and describe a broader spectral range to relate to.

The advantage of these figures is that they dichotomize the 12 room pairs into 2 very distinct groups, something that does not happen with any other alternative for difference calculation or frequency band combinations so clearly. Figures 7.4 and 7.5 below show the simple logistic model that was computed using BTD as the only independent variable in the logistic regression in parts 1 and 2 separately. The goodness-of-fit is somewhat different for the two parts, better for the first than the second and best for both combined, (figure 7.6).

The sum of squared residuals from each logistic regression should be minimal for the best fit. Using T_{30} with BTD has the smallest sums of all three parameters. It is probable that the measured values of the T_{30} are the most accurate, as was explained in section 3.1.3 and therefore produce the most reliable parameter. The resultant logistic regressions using BED and BTD2 are slightly less successful and are not shown here.

As the effect of BTD was so dominant only one more independent variable was examined when added to the logistic regression, which was the program material. It is assigned as a categorical binary variable, which equals 1 when speech and 0 otherwise. The resultant curve is shown in figure 7.6. Further testing into different musical programs was not examined. Although statistically the goodness-of-fit is only a bit better than before, it does show an interesting feature, by raising the entire curve for speech pairs. That, in turn, predicts higher probabilities to match a pair even at smaller BTD, which corroborates examinees' reports for relative easiness in matching speech pairs. Table 7.4 summarizes some of the quantities mentioned for all items in experiment B.

However, interpretation of the BTD quantity, in spite of the above, should be done cautiously for several reasons, as this strong association does not necessarily point to causation. The only thing it

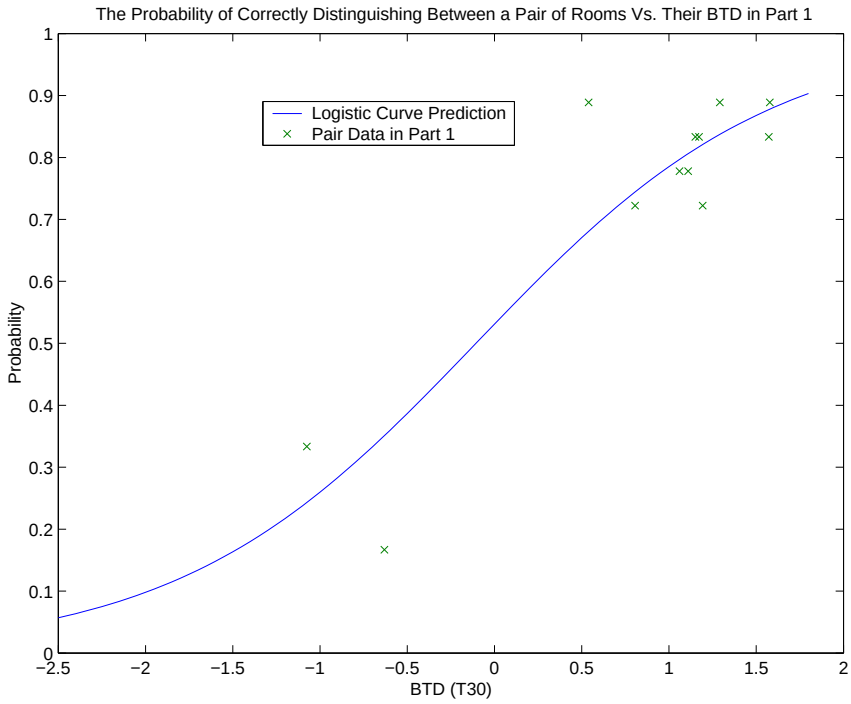


Figure 7.4: Logistic regression prediction of the probability of correctly answering test items in part 1 of Experiment B, as a function of the room pair Bass T_{30} Difference. The sum of squared deviance residuals from the logistic regression is 10.1074.

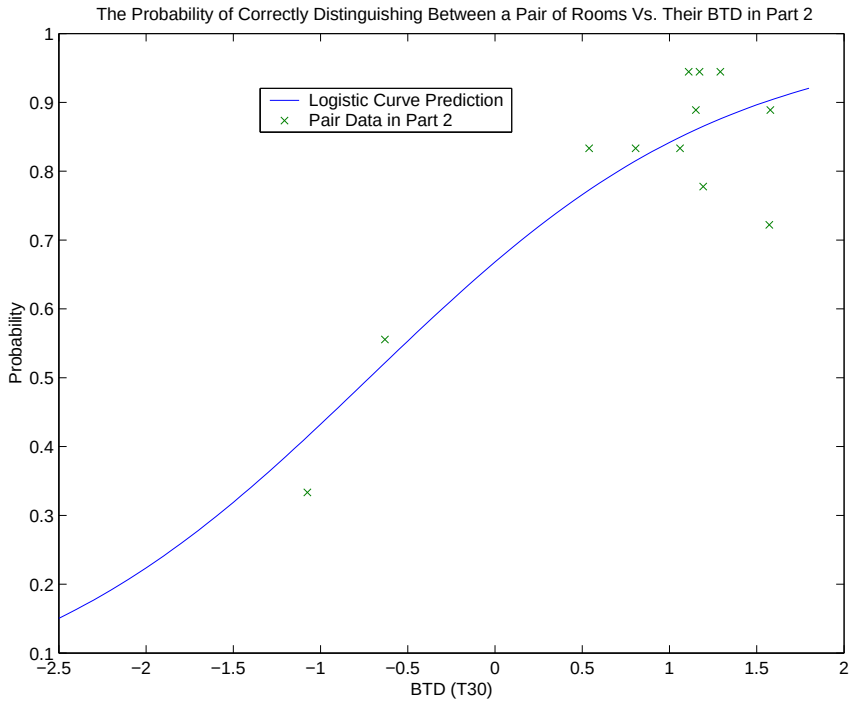


Figure 7.5: Logistic regression prediction of the probability of correctly answering test items in part 2 of Experiment B, as a function of the room pair Bass T_{30} Difference. The sum of squared deviance residuals from the logistic regression is 10.6069

suggests is that, on average, pairs of small rooms with a small values of BTD, BED or BTD2 are likely to sound confusingly similar. It also means that subjects who fared better, must have based their decisions on more than just the low RT criteria. Finally, the most surprising aspect, which underlines the necessity of validating the results through more experiments, is the relatively low frequency bands encompassed by the BTD, which shed some light over a frequency range which is often left out or is rather given a large tolerance in design. Together the three quantities cover a range frequency range between $44 - 180Hz$ (taking into account the full width of the $50 - 160Hz$ one-third octave bands). Before drawing further conclusions, let us look at Experiment C.

Experiment C

The task in Experiment C, although still defined the same as in Experiment B, was presented as, and proved to be, much harder. It is likely that a few subjects guessed in many cases and there are only 3 test items, which on the whole show significantly higher probability levels than chance. These levels are still lower compared to the previous experiment, even in items with the same 2 room pairs. In fact, all questions in the test, but one, have their equivalent in Experiment B, save for the programs and positions in the room. Therefore, a reasonable premise would say that a difficult item in Experiment B cannot be easy in Experiment C. There is only one such common item - question 2 in both tests, which compares the DR control studio with the talk studio. It also implies that the quantities previously described that may have been used to determine correct pairing is no longer sufficient to account for the results here. Applying BDT or BET would mean that only question 2 is difficult, which is clearly wrong.

However, reviewing the results and averages over all subjects and items, no consistency was found between quantities which could distinguish between difficulty levels, using one independent variable only. Various combinations of RT-derived quantities, room volume and even room boxiness from experiment A were tested to no avail. And yet there were three significantly likely test items to be correctly answered. Another close inspection on the sets of pairs and their programs, revealed 4 items, in which the division between the 4 programs happened to be the same: one room having “Male Speech” and “Jimi” and the other “Mingus” and “Herbert”. On the face on it, it should be no different than other items, as the same programs are used, but all of these items had very low scores. A categorical variable was set, which equals 1 in items which have the same order and 0 otherwise. Together with the BTD (defined the same as in Experiment B) a logistic regression model was fitted, shown in figures 7.7, 7.8 and 7.9.

This above model is hardly satisfying. The underlying reason behind this bias is unclear as subjects deal with the same samples throughout the entire test. If it is indeed valid it means that an average examinee, who bases his decision on something similar the BTD, is being led into guessing with one particular division or programs between the rooms. In turn it suggests that there is a tremendous subjective difficulty to extract this information from four different programs, insofar as that it depends on program division between rooms. Finally, it reinforces the preliminary observation that showed that subjects who fared well in one experiment, did not necessarily manage better in the second. The same useful criterion for Experiment B is much harder to apply in Experiment C.

All that given, another approach was attempted. The 5 best subjects in Experiment C, ranked by their overall scores, were treated as an “expert team”, which was analyzed separately. Immediate inspection of the new item probabilities revealed another easy item and generally a better dichotomy between easy and difficult items. The search for a single parameter was started anew and soon converged. The single independent variable, which is referred to here as MTD – Midrange T_{30} Difference – the difference between the $T30_{30}$ means in the frequencies bands $250 - 2500Hz$. As before, a similar quantity can be derived from the EDT, only this time not as good as the one from T_{30}. The optimized definitions used are:

$$MTD = \left| \frac{1}{N} \sum_{f=250Hz}^{2500Hz} T_{30,Room1}(f) - \frac{1}{N} \sum_{f=250Hz}^{2500Hz} T_{30,Room2}(f) \right| = \left| \frac{1}{N} \sum_{f=250Hz}^{2500Hz} (T_{30,Room1}(f) - T_{30,Room2}(f)) \right| \tag{7.13}$$

Item B	Item C	$p_{t,B}$	BTD	BED	$p_{t,C}$	$p_{t,Ce}$	MTD	MED	Rooms
B1	C3	0.8611	1.1099	0.8616	0.7778	0.9000	0.2450	0.1900	L,B
B2	C2	0.3333	-1.0759	-1.3704	0.2500	0.3000	0.0718	0.0640	C,T
B3	-	0.8611	0.5394	0.4600	-	-	0.2042	0.1322	M,B
B4(S)	C4	0.8889	1.1719	0.7962	0.5556	0.7000	0.2360	0.2094	T,I
B5	-	0.7778	1.5715	1.4071	-	-	0.2875	0.2047	M,C
B6(S)	C6(P)	0.9167	1.2914	1.0654	0.1667	0.2000	0.1551	0.1364	T,B
B7	-	0.8611	1.1512	0.8604	-	-	0.3513	0.3484	M,H
B8	C8(P)	0.8056	1.0602	0.7934	0.1389	0.2000	0.1642	0.1454	C,I
B9(S)	C7(P)	0.8889	1.5779	1.3279	0.3056	0.9000	0.3105	0.2907	L,H
B10	-	0.7500	1.1930	1.0633	-	-	0.0833	0.0724	C,B
B11	C1(P)	0.7778	0.8060	0.6876	0.2778	0.3000	0.0408	0.0578	M,L
B12	-	0.3611	-0.6311	-0.3783	-	-	0.0809	0.0730	B,I
-	C5	-	1.6400	1.4085	0.7222	1.0000	0.3593	0.2687	M,B

Table 7.4: Summary of item probability in both parts in experiments B and C $p_{t,B}$ and $p_{t,C}$, BTD, BED, MTD, MED, probability for the expert group $p_{t,Ce}$ and the pair of rooms, where the following nomenclature is used: B - Library, C - Control Room 3, H - Hearing Protector Testing Room, I - IEC Listening Room, L - Lecture Room, M - Meeting Room, T - Talk Studio 8. Items marked with (S) were tested for speech probability improvement in Exp. B and items with (P) for program division bias in Exp. C.

$$MED = \left| \frac{1}{N} \sum_{f=315Hz}^{2000Hz} EDT_{Room1}(f) - \frac{1}{N} \sum_{f=315Hz}^{2000Hz} EDT_{Room2}(f) \right| = \left| \frac{1}{N} \sum_{f=315Hz}^{2000Hz} (EDT_{Room1}(f) - EDT_{Room2}(f)) \right| \tag{7.14}$$

Where N are the number of terms in the summation of the one-third octave band values.

Since the model predictions are based on very few subjects they are very coarse. It should be noted that the frequency limiting bands are not well-defined and adding or subtracting a few terms does not affect dramatically the fit. Also, there seems to be a very narrow threshold range, which dichotomizes difficult and easy items. It is especially narrow with MED and less so with MTD. Values of most quantities used for the analysis of all items in experiment C are brought in table 7.4 side by side with the equivalent items from experiment B.

It is possible to speculate on the relation between the RT midrange curve at the MTD frequencies, whether they convey information about the room shape and dimensions. If indeed subjects relied on their focus to hear to room size, shape, walls, volume and such (in their words) then maybe this capability is related more directly to the RT of the room.

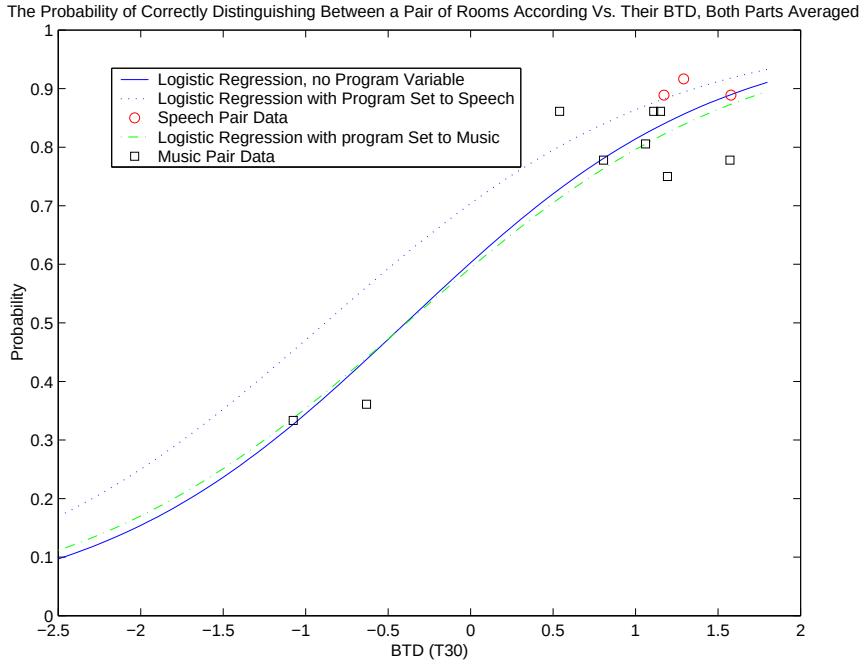


Figure 7.6: Logistic regression prediction of the probability of correctly answering items in both parts of Experiment B, as a function of the room pair Bass T_{30} Difference and test program. The sum of squared deviance residuals from the logistic regression is 12.6326 for the single variable and 10.7912 for the two-variable fit.

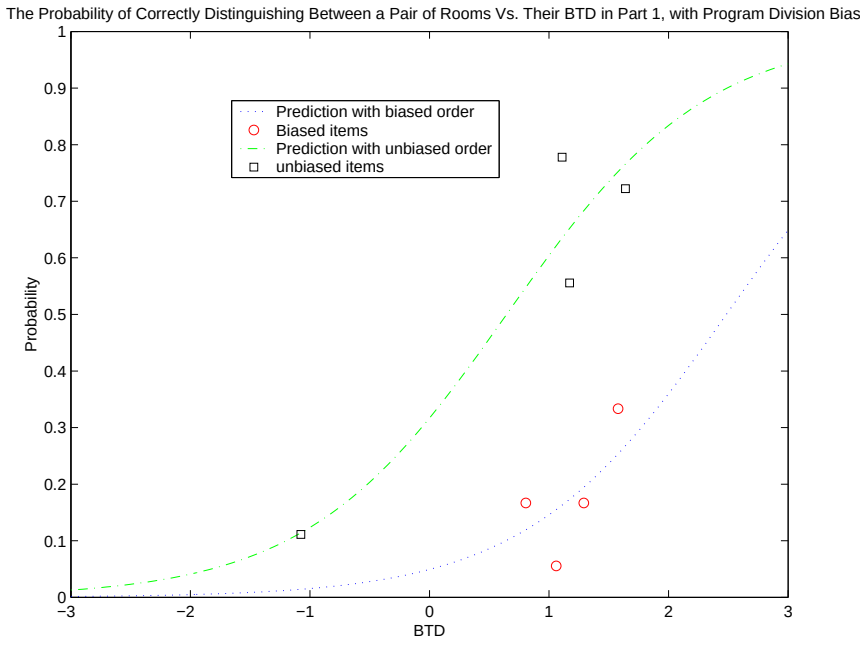


Figure 7.7: Logistic regression prediction of the probability of correctly answering items in part 1 of Experiment C, as a function of the room pair Bass T_{30} Difference and possible bias when a particular program division between the two rooms was used.

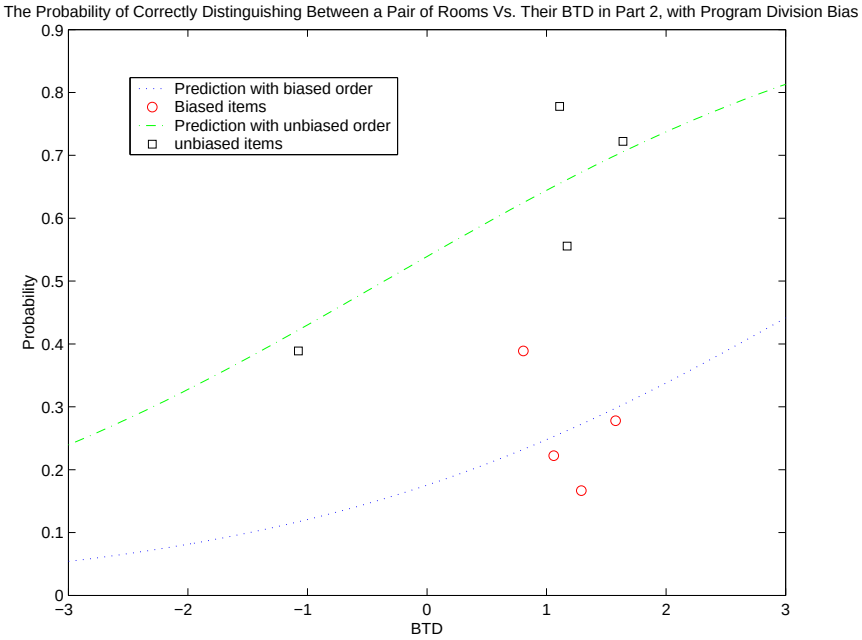


Figure 7.8: Logistic regression prediction of the probability of correctly answering test in part 2 of Experiment C, as a function of the room pair Bass T_{30} Difference and possible bias when a particular program division between the two rooms was used..

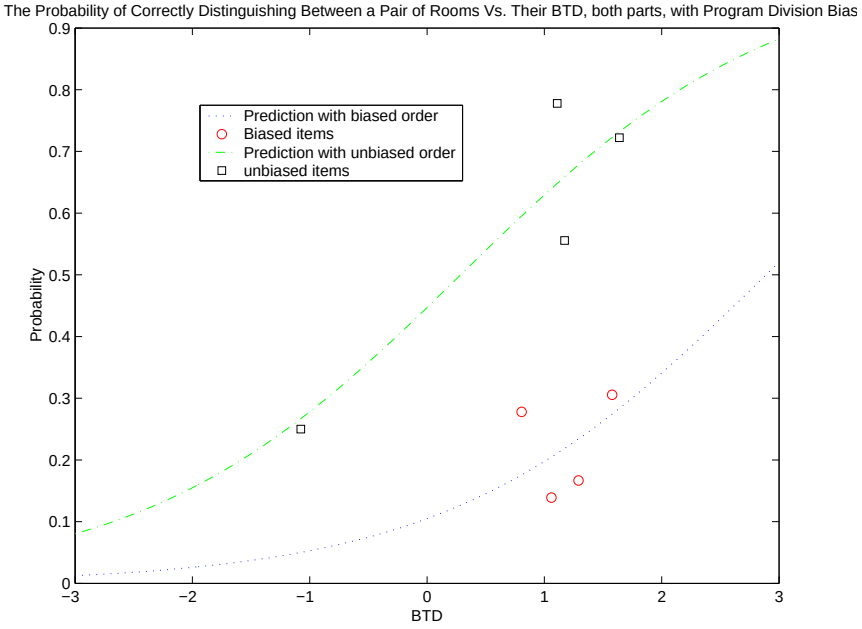


Figure 7.9: Logistic regression prediction of the probability of correctly answering test in both parts of Experiment C, as a function of the room pair Bass T_{30} Difference and possible bias when a particular program division between the two rooms was used.

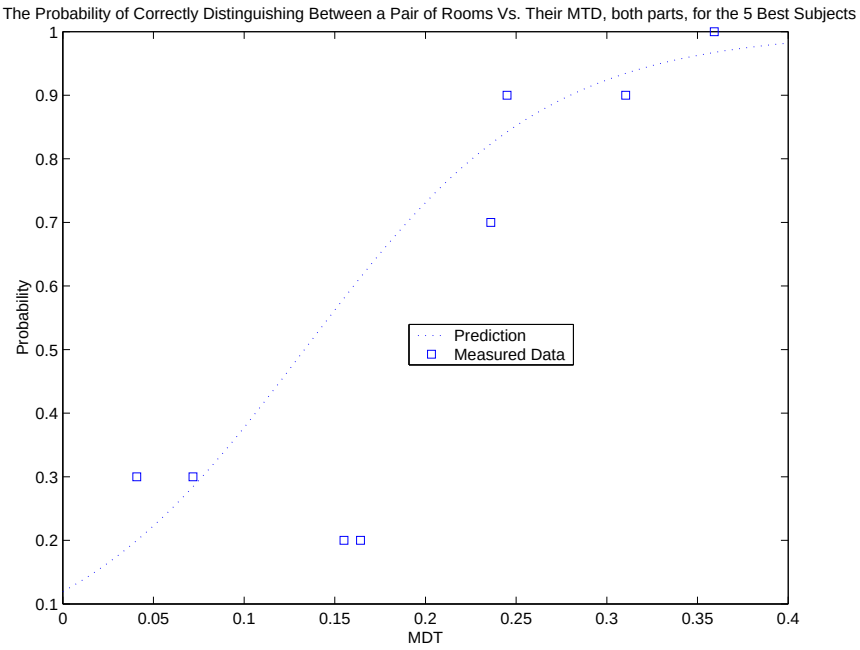


Figure 7.10: Logistic regression prediction of the probability of correctly answering test items for the 5 best subjects in both parts of Experiment C, as a function of the room pair mean midrange T_{30} difference.

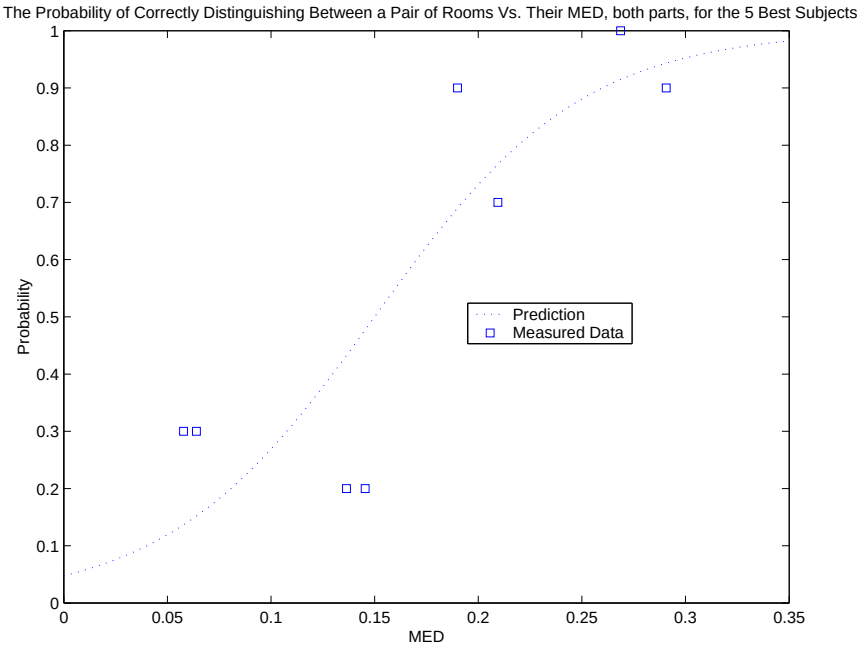


Figure 7.11: Logistic regression prediction of the probability of correctly answering test items for the 5 best subjects in both parts of Experiment C, as a function of the room pair mean midrange EDT difference.

Chapter 8

Discussion and Conclusions

This discussion covers the following subjects:

- Project premises and validity
- Experiment A
- Experiments B and C
- General conclusions
- Addressing the preliminary research questions

8.1 Project Premises and Validity

Three novel experiments were introduced in this research. All three make use of an extensive binaurally recorded library, which contains 230 recordings of 10 different programs, recorded in 7 different small rooms, in total of 23 source-receiver recording positions. The source used was a monophonic loudspeaker and recorded binaurally with a head and torso simulator, both of the highest quality attainable. The major acoustical parameters of the rooms were measured in conjunction with the recordings. A dedicated computerized interface was developed for subsequent automated subjective tests. Eighteen test subjects performed the three tests. Each test was repeated twice.

The small rooms measured were of volumes between $26m^3$ and $187m^3$. Mean reverberation times varied between $0.12s$ to $0.83s$ in the $200 - 4000Hz$ range.

The fundamental premise, on which all three experiments rely, is that the binaural recordings representation, employed to simulate the various room acoustics, is a reliable technique, which is telling of the original acoustics it replicates to a high degree of accuracy and authenticity. In other words, the basic acoustical parameters of the rooms, most notably the reverberation time, are perceptually invariable through the binaural representation.

Another premise is that all the contributions to the recordings from equipment imperfections, such as harmonic distortion, noise and limited dynamic range, form a constant distortion, which is also invariable between representations.

In the strictest sense the test subjects in the experiment are not a representative sampling of the population, as most are acousticians or other students. One should be cautious before generalizing the results.

The final frequency bands that were used for calculation in all the new quantities in all three experiments should be treated with caution too. The particular bands may have been correct for these particular cases, but might prove ineffective in other setups. It is more reasonable to confide in the more general

functional relations that describe these quantities, which can be optimized using different combinations of frequency bands in the vicinity the bands obtained here.

In view of the above, the exact validity of the tests is unknown. Seldom could the findings be compared to the literature to validate their results. Therefore, further research is needed in order to be able to compare findings obtained in different methods. A few rough ideas are suggested along the discussion.

8.2 Experiment A

Experiment A tested subject preferences in terms of what was rather vaguely called "overall sound quality". The programs tested were a male anechoic speech and a studio recording of an electronic popular music track. A representative source-position combination was chosen of each room. Two more parameters were added to subjective evaluation - "boxiness" and "boominess" - both in attempt to separate later what could account as adverse effects on the overall sound quality.

Bass ratio in bigger halls is not necessarily associated with preferable sound quality, but this concept seems handy as an instrument to describe the deviations from a flat reverberation time curve, which is of main interest here. Three new ratios are introduced, which all use only one-third octave band values and span on two-thirds of an octave only in both their numerators and denominators.

8.2.1 Summary of Main Findings

The main findings of the experiment are:

- Ratings usually showed different preferences for the speech and the musical programs in terms of the overall sound quality of the recordings.
- Classical bass ratio definitions showed only poor correlations with all ratings and new quantities had to be sought.
- Speech sound quality was generally rated higher for rooms with lower mean RT or EDT ($200 - 4000Hz$).
- 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 describes the difference between the high treble to the low bass RT in the rooms.
- Music SQ ratings peaked at higher mean RT's (and EDT) between $0.3s$ and $0.5s$. It also showed the highest correlation with the Small room Bass Ratio (SBR). However the music SQ could not be modeled using only the reviewed and rated parameters and it is likely that there is at least one more hidden parameter which was not measured, which is instrumental to understanding of the perceived SQ for music.
- When music and speech are combined to one total SQ rating it displays high correlations with both SBR and SEBR.
- Boxiness showed, in a way, less dependency on the program material. It inversely correlated well with SEBR in all cases and rather well with T_{30} and EDT.
- Boominess could only be correlated, to some degree, with the LHR, especially in the music rating.
- The mean RT, as recommended by the listening room standards using the room volume, showed also some correlation with both sound quality and inversely with boxiness.

Findings regarding specific rooms:

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

- The poorest two rooms are the lecture room and the even worse 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 single modes in the room, which encounter no damping mechanism.

8.2.2 Conclusions and Further Research Paths

The three new quantities SBR, SEBR and LHR show the strongest relations to the SQ, boxiness and boominess in the small rooms measured compared to classical parameters. Whether they are valid measures of other small rooms with different program material and perhaps different loudspeaker used, is a subject for further research. It is believed 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, then it shows increased subject sensitivity to very low frequency bands ranging from 50 to 100Hz. Results from Experiments B and C reinforce these findings. The reason for difference in correlations and performance by using the SBR or the SEBR is unclear.

The ratings of boominess and boxiness did not fulfill the original intention. Boxiness ratings showed high dependency 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 non-linearly. There is a doubt of how well the subjects understood both terms (logical error).

A simple survey can be made by using existing measured RT curves of various rooms and calculating these parameters for them. Thereafter, they can 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 in the fashion of the one by Gade [51] and [52] about musicians' conditions, can then account for all the factors the define the overall sound quality in small rooms. An important subjective parameter, which was left out here, is timbre.

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

8.3 Experiments B and C

Experiment B tested the capability of listeners to match a pair of recordings made in the same room, but in different source-receiver positions, by distinguishing the matching pair and two other recordings of a different room pair. Twelve room pairs were presented of four programs, each used for three questions.

The main findings of Experiment B are:

- A single defining parameter was sufficient to explain the results in the test. According to this parameter, subjects were able to more easily discern RT differences between rooms.

- Alternative RT-difference parameters, which were tested, eventually converged to a narrow frequency range at very low one-third octave bands between 50-160Hz.

Experiment C is a variation on Experiment B, in which the four samples presented are of different programs, still recorded in two different rooms. The task is still to tell which two samples were recorded in the same room. Eight questions were presented.

The main findings of Experiment C are:

- On the whole, the performance in this experiment was very poor compared to Experiment B.
- The single parameter from Experiment B could not be used to convincingly predict the results here, unless an order bias (sequential error) was introduced into the prediction model.
- A small ("expert") group of subjects was still able to correctly match pairs, despite the enhanced difficulty of the test.
- The expert group scores can be related to RT differences between the rooms in the $250 - 3150Hz$ frequency range measured in one-third octave bands.

8.3.1 Conclusions

Differences between predictions based are not clear. The variations in the goodness of fit shown in parameters derived from T_{30}, T_{20} and EDT may well be a subject to uncertainties in the RT measurements or sampling error of the test data. Despite the $1\frac{2}{3}$ octaves span of the difference in used for the Bass T_{30} Difference (BTD), the $50Hz$ one-third octave band appears to have a detrimental role in the parameter, as a single band.

In informal acoustical-ecological terms, both experiments point to the human capability to hear through complex signals and, in many cases, correctly detect the environment in which they were recorded. It must be said that the test is artificial in nature: when people hear a signal in various points within a space, they move around continuously and do not leap from one point to another. The continuity in real life deemphasizes the variations in sound that are only accentuated in these experiments. In a way it draws a parallel between the exact reverberation time in the room and its auditory imprint. Although the reasoning for that can be related to single defining RT-derived quantities, their broadband represented in the two experiments suggests a more comprehensive integrating listening, which is not confined to narrow definitions like these.

It should be noted that no circumstantial evidence was found that supports the viability of room volume perception, as was initially conjectured and thought to be supported by results from [53] and [54]. Although volume perception is likely to be highly affected by the degree of success to which the binaural representation works, it seems that the room reverberance often misleads the listener and creates a false room volume, when no visual information is available.

Similar tests can be made, which are considerably more difficult - comparing more than two rooms at once, having with more alternatives, etc.

8.4 General Conclusions

All experiments rely heavily on the reverberation time in the small rooms. All the new quantities are derived directly from the RT curves and are generalized to describe many of the observations in the experiments, without resorting to more complicated room parameters.

Moreover, in both Experiment A and B, the findings show how sensitive listeners are to very low frequencies, down to the $50Hz$ one-third octave band.

No direct connection was found between the tests. It was examined whether room ratings in Experiment A that were not significantly different than each other, could be more difficult to distinguish in the pair matching tests. This was not the case. Therefore, a preferable sound does not necessarily entail an easily detectable single quality or quantity that can be pinned down and related to its original room acoustics.

All experiments tell us something about the "participation" of the room acoustics in recordings. It is well-known that the room has a crucial effect on recordings and on critical listening. However, whether preferred rooms are in a sense more neutral, or less dominant, than others in their imprint on the recordings, is left to further investigation.

8.5 Addressing the Preliminary Research Questions

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.

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

The non-flat reverberation time curve issue can be addressed to some extent. Repeating the introductory question: is a longer (on average) yet flat RT curve preferable over 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) vs. longer non-flat RT curve (library, IEC listening room and even the meeting room). Subjects preferred the more reverberant bass of the latter group over the more dry former group in the music ratings. In speech the results were mixed, but the library and talk studio are comparably good. The combined ratings show clear preference to the more reverberant room group.

The exact interrelation between coloration and reverberation time at low frequencies was left to future exploration.

No real critique can be contested regarding the listening room standards, as no measured room complied with all the specifications for a standardized room. Nevertheless, the library adhered closest to the standard requirements and indeed scored best.

Appendix A

List of Abbreviations

BED - Bass EDT (q.v.) Difference
BR - Bass Ratio
BTD - Bass T_{30} Difference
BTD2 - Bass T_{20} Difference
EDT - Early Decay Time
FFT - Fast Fourier Transform
FHT - Fast Hadamard Transform
HATS - Head And Torso Simulator
HDR - Hard Disk Recorder
HRTF - Head Related Transfer Function
HpTF - Headphone Transfer Function
INR - Impulse Response to Noise Ratio
LHR - Low High Ratio
MLS - Maximum Length Sequence
MTD - Midrange T_{30} Difference
PTF - Headphone Transfer Function
RT - Reverberation Time
SBR - Small room Bass Ratio
SEBR - Small room EDT (q.v.) Bass Ratio
SPL - Sound Pressure Level
SQ - Sound Quality

Appendix B

Matlab Codes for the Subjective Tests

Three Matlab codes were written especially for the project. They were tailored to sample library under test and to other special requirements, such as output and input texts. However, if the need arises, modifying them into other similar tests should not pose a big problem for a person who is familiar with basic Matlab programming.

B.1 Latin Square Generation with "Latin"

This simple routine generates an even numbered latin square sequence of up to 20 elements, upon a specified input of the square order. It randomly picks a line out of the square and saves a file of the lines that have been already tested for this particular square. When the number of already tested sequences equals the square's order, the file is reset. The function returns a row vector of the square line as output.

All other codes use *Latin* for randomization.

```
%Even Latin Square Randomization Routine
function [order]=latin(n) %returns a latin sqaure row for a given number of terms (even only)
latin_series=[1 2 n 3 (n-1) 4 (n-2) 5 (n-3) 6 (n-4) 7 (n-5) 8 (n-6) 9 (n-7) 10 (n-8) 11]; %Up to 20 terms in the series
try % Make sure that the last experiment orders are not repeated
    N=dlmread(['already_done' num2str(n) '.txt'],'\t'); %Reads the file of the recent order
catch % if doesn't exist
    N=-1;
end
if length(N)==n %when reached one cycle of all rows - start over
    N=-1;
end
k=unidrnd(n)-1; % Choose a random number of row
t=1;
while t<=length(N) % check if it was tested in the recent cycle
    if k == N(t)
        k=unidrnd(n)-1;
        t=0;
    end
    t=t+1;
end
if N(1)==-1 % If the cycle has just begun
    N(1)=k; % create the first term in file
else
    N(length(N)+1)=k; % Or add an additional term
end
```

```
dlmwrite(['already_done' num2str(n) '.txt'],N,'\t');
for t=1:n
order(t)=mod(latin_series(t)+k,n); %create the actual latin square row
if order(t) ==0
    order(t)=n;
end
end
```

B.2 Experiment A - "expa"

Much of this code is self-explanatory with the adjoined remarks and the texts. It reads the files in a predefined directory, which all have the same naming convention. The same code is basically repeated twice for the examples and for the test itself.

It generates 4 output files, named after the subject's name: summary of all scores in the test, a matrix of the sound quality ratings, matrix for boxiness and a matrix for boominess.

```
runs=14; %Number of tracks in each repetition
parts=2; %Number of repetitions
clc
Herbert=dir('C:\Documents and Settings\AWat\My Documents\Listening Tests\ExpA Tracks\Herbert*');
Male=dir('C:\Documents and Settings\AWat\My Documents\Listening Tests\ExpA Tracks\Male*');
All={Herbert(1:7).name Male(1:7).name}; %form a cell array with all the file names to be played
All=char(All); %Convert to a strings
n=size(All,1); %Number of files
curdir='C:\Documents and Settings\AWat\My Documents\Listening Tests\ExpA Tracks\';
example={'C:\Documents and Settings\AWat\My Documents\Listening Tests\ExpA Tracks\Jimi - DR Talk Studio 8 - 3.wav'
 'C:\Documents and Settings\AWat\My Documents\Listening Tests\ExpA Tracks\Jimi - HP Room 3.wav'
 'C:\Documents and Settings\AWat\My Documents\Listening Tests\ExpA Tracks\Jimi - Lecture Room 3.wav'};
%clc
skip_flag=input('Press E and ENTER to go through the introduction or just ENTER to skip it ','s');
warning off;
if (skip_flag=='E') | (skip_flag=='e')
    clc
    fprintf('\nINTRODUCTORY EXAMPLES\n You are going to be presented with a series of recorded samples.
     Each one was recorded in a different room and is 15-20 seconds long.\n')
    fprintf('\nPlease listen carefully to the samples and their overall sound quality. Is it a good recording?
     Is it boomy? And does it sound ''boxy''?\n')
    fprintf('In the following examples you may practice your scaling for each of the above three aspects.
     \nScale each parameter independently of the other two.\n\n')
    for t = 1:3 %number of examples
        sample_name=char(example(t));%splice the entire file name; uses the first 3 files in the directory
        Sam_Size=WAVREAD(sample_name,'size');
        [Y,FS,NBITS]=WAVREAD(sample_name,Sam_Size(1));% read first 1 second
        fprintf('\n SAMPLE %d \n',t)
        fprintf('\n Listen carefully to the overall sound quality of the following sample. Also note its ''boominess
         and its ''boxiness''.\n Please ignore any left-right balance and volume artifacts of the recordings.\n')
        fprintf('\n Press any key when ready to hear the sample\n')
        pause
        soundsc(Y,FS)
        pause(Sam_Size(1)/44100)
        s=input('\nPress A and ENTER to listen to the sample again or just ENTER to continue ','s');
        while (s=='A') | (s=='a')
            sound(Y,FS)
            pause(Sam_Size(1)/44100)
            s='';
            s=input('\nPress A and ENTER to listen to the sample again or just ENTER to continue ','s');
        end
        clc
        fprintf('\nOn a 1-9 scale, 1 being the poorest quality and 9 the best quality, how would you rate the
         overall sound quality of the sample?');
        a=input('\nUse the scale below:\n1-intolerable\n2-Very Annoying\n3-Unpleasant\n4-Not So Good\n5-Acceptable
```

```
         \n6-Decent\n7-Good\n8-Very Good\n9-Excellent\n',’s’);
        a=str2num(a);
        while (isempty(a) | (a<1 | a>9))
            a=input(’\n\n Please enter a valid grade between 1 and 9: ’,’s’);
            a=str2num(a);
        end
        clc
        fprintf(’\nOn a 1-7 scale, how boomy is the sample?’);
        b=input(’\nUse the scale below:\n1-Very Thin, Very Hollow\n2-Thin, Hollow\n3-Slightly Thin, Hollow\n4-Balanced
         \n5-Slightly Boomy\n6-Boomy\n7-Very Emphasized\n’,’s’);
        b=str2num(b);
        while (isempty(b) | (b<1 | b>7))
            b=input(’\n\n Please enter a valid grade between 1 and 7: ’,’s’);
            b=str2num(b);
        end
        clc
        fprintf(’\nOn a 1-5 scale, how boxy is the sample?’);
        c=input(’\nUse the scale below:\n1-Unnoticeable\n2-Barely Audible\n3-Distinct yet not dominant\n4-Dominant
         \n5-Very Dominant\n’,’s’);
        c=str2num(c);
        while (isempty(c) | (c<1 | c>5))
            c=input(’\n\n Please enter a valid grade between 1 and 5: ’,’s’);
            c=str2num(c);
        end
        clc
    end
    fprintf(’ Introduction over.\n\n Press any key to begin with the experiment.’)
    pause
end
subject=input(’\nPlease enter your name: ’,’s’);
fid=fopen(strcat(’Experiment_A_score_’,subject,’.txt’),’w’); %Creates a file for the subject
fprintf(fid,’%s \t\t %s \r\n’,subject, date);
fprintf(fid,’Filename %s \t Sound Quality \t Boominess \t Boxiness \r\n\r\n’,blanks(24));
SQ=zeros(size(All,1),parts); %Empty sound quality matrix
Boominess=zeros(size(All,1),parts); %Empty boominess matrix
Boxiness=zeros(size(All,1),parts); %Empty boxiness matrix
clc
for r= 1:parts
    All_order=latin(n); % call a latin square row for each part. assuming an even repetition number, it is assured
     that they will not be identical.
    fprintf(3,’PART %d \r\n’,r);
    for t = 1:runs
        clc
        sample_name=strcat(’C:\Documents and Settings\AWat\My Documents\Listening Tests\ExpA Tracks\’,
         All(All_order(t),:));%splice the entire file name
        Sam_Size=WAVREAD(sample_name,’size’);
        [Y,FS,NBITS]=WAVREAD(sample_name,Sam_Size(1));% read first 1 second
        fprintf(’\n SAMPLE %d \n’,t)
        fprintf(’\n Listen carefully to the overall sound quality of the following sample. Also note its
         ’’boominess’’ and its ’’boxiness’’. \n Please ignore any left-right balance and volume artifacts
          of the recordings.\n’)
        fprintf(’\n Press any key when ready to hear the sample\n’)
        pause
        sound(Y,FS)
        pause(Sam_Size(1)/44100)
        s=input(’\nPress A and ENTER to listen to the sample again or just ENTER key to continue ’,’s’);
        while (s==’A’) | (s==’a’)
            sound(Y,FS)
            pause(Sam_Size(1)/44100)
            s=’’;
            s=input(’\nPress A and ENTER to listen to the sample again or just ENTER to continue ’,’s’);
        end
        clc
```

```
        fprintf('\nOn a 1-9 scale, 1 being the poorest quality and 9 the best quality, how would you rate the
         overall sound quality of the sample?');
        a=input('\nUse the scale below:\n1-intolerable\n2-Very Annoying\n3-Unpleasant\n4-Not So Good
         \n5-Acceptable\n6-Decent\n7-Good\n8-Very Good\n9-Excellent\n','s');
        a=str2num(a);
        while (isempty(a) | (a<1 | a>9))
            a=input('\n\n Please enter a valid grade between 1 and 9: ','s');
            a=str2num(a);
        end
        SQ(All_order(t),r)=a;
        clc
        fprintf('\nOn a 1-7 scale, how boomy is the sample?');
        b=input('\nUse the scale below:\n1-Very Thin, Very Hollow\n2-Thin, Hollow\n3-Slightly Thin, Hollow
         \n4-Balanced\n5-Slightly Boomy\n6-Boomy\n7-Very Emphasized\n','s');
        b=str2num(b);
        while (isempty(b) | (b<1 | b>7))
            b=input('\n\n Please enter a valid grade between 1 and 7: ','s');
            b=str2num(b);
        end
        Boominess(All_order(t),r)=b;
        clc
        fprintf('\nOn a 1-5 scale, how boxy is the sample?');
        c=input('\nUse the scale below:\n1-Unnoticeable\n2-Barely Audible\n3-Distinct yet not dominant
         \n4-Dominant\n5-Very Dominant\n','s');
        c=str2num(c);
        while (isempty(c) | (c<1 | c>5))
            c=input('\n\n Please enter a valid grade between 1 and 5: ','s');
            c=str2num(c);
        end
        Boxiness(All_order(t),r)=c;
    end
    for t=1:n
        fprintf(fid,'%s \t %d \t %d \t %d \r\n',[All(t,:) blanks(30-length(All(t,:)))],SQ(t,r),
         Boominess(t,r),Boxiness(t,r)); %writes the results to file
    end
    if r ~= parts
        clc
        fprintf('End of part %d.\n\n Please take a few minutes break and when ready press any key to continue.\n',r)
        pause
    end
end
clc
fprintf('Thank you!\n')
fclose(fid);
%export matrices
dlmwrite(strcat('Experiment_A_SQ_',subject,'.txt'),SQ,'\t');
dlmwrite(strcat('Experiment_A_Boominess_',subject,'.txt'),Boominess,'\t');
dlmwrite(strcat('Experiment_A_Boxiness_',subject,'.txt'),Boxiness,'\t');
```

B.3 Window Interface for Experiments B and C - "Bico"

Due to unsolved bugs in the Matlab audioplayer function, alternative commands had to be sought. The snd_pc commands have to be downloaded and installed on the according to the instructions in the readme file in order to run the following codes. [1]

"Bico" is a special GUI file written using Matlab's GUIDE. It cannot be brought here though. All commands are inputed through the so-called property manager of the various objects in the window. See figure 5.1 for the window illustration.

[1] Available for download from
www.mathworks.com/matlabcentral/fileexchange/download.do?objectId=112&fn=snd_pc&fe=.zip&cid=803729

Both "expb" and "expc" call "Bico" for each test item.

B.4 Experiments B and C - "expb" and "expc"

In these codes, a special file naming scheme is used in addition to assigning them into separate folders. Folders 1 and 2 for one room and 3 and 4 for another. The same file number would then represent that the 4 files make two pairs, or a test item. This rather cumbersome system was created to form complete randomization of the presented samples under the A, B, C or D buttons in the "Bico" window.

The code generates three files under the subject's name: a summary of all scores in both parts, a vector of the scores in part 1, score vector for part 2.

"expc" is practically identical to "expb", apart from a few different path names and the total item number and is not reprinted here.

```
curdir='C:\Documents and Settings\AWat\My Documents\Listening Tests\';
parts=2;
rand('state',sum(100*clock));
clc
skip_flag=input('Press E and ENTER to go through the introduction or just ENTER to skip it ','s');
warning off;
if (skip_flag=='E') | (skip_flag=='e')
    clc
    fprintf('\n\nINTRODUCTORY EXAMPLES\n\n In the following tests you are asked to recognize two out of four
      given recorded samples A, B, C and D that were recorded \nin the same room, but at different positions. ')
    fprintf('The two other samples were recorded in another room. \nListen closely to each and once you have decided,
     use the mouse to press either button\n ''Rooms A and B are the same'', ''Rooms A and C are the same'' or
      ''Rooms A and D are the same''.\n')
    fprintf('\nPlease ignore any L-R channel balance and recording volume differences between samples, as they are
     irrelevant to the particular task.\n')
    fprintf('\nHere are a couple of examples for familiarization, both with the questions and the degree
     of \ndifference between recordings you might have to deal with.\n')
    fprintf('Press any key to continue.')
    pause
    clc
    exampleA={'Female - DR Talk Studio 8 - 3' 'The - IEC Room 2'};
    exampleB={'Female - HP Room 1' 'The - Library 1'};
    exampleC={'Female - HP Room 3' 'The - IEC Room 4'};
    exampleD={'Female - DR Talk Studio 8 - 2' 'The - Library 3'};
    for t = 1:2
        hit=0;
        clc
        sample_nameA=char(strcat(curdir,'examples\',exampleA(t),'.wav'));
        sample_nameB=char(strcat(curdir,'examples\',exampleB(t),'.wav'));
        sample_nameC=char(strcat(curdir,'examples\',exampleC(t),'.wav'));
        sample_nameD=char(strcat(curdir,'examples\',exampleD(t),'.wav'));
        SizeA=WAVREAD(sample_nameA,'size');
        SizeB=WAVREAD(sample_nameB,'size');
        SizeC=WAVREAD(sample_nameC,'size');
        SizeD=WAVREAD(sample_nameD,'size');
        [YA,FS,NBITS]=WAVREAD(sample_nameA,SizeA(1));
        [YB,FS,NBITS]=WAVREAD(sample_nameB,SizeB(1));
        [YC,FS,NBITS]=WAVREAD(sample_nameC,SizeC(1));
        [YD,FS,NBITS]=WAVREAD(sample_nameD,SizeD(1));
        pA = SND_MULTI([1 0 44100 16],ones(0,100));
        pB = SND_MULTI([1 0 44100 16],ones(0,100));
        pC = SND_MULTI([1 0 44100 16],ones(0,100));
        pD = SND_MULTI([1 0 44100 16],ones(0,100));
        Bico
        if t==1
            if hit==3
                fprintf(' Correct!')
```

```
            else
                fprintf(' Wrong answer. The right answer was ''Rooms A and D are the same''.\nPress any key to
                 see that example again.')
                pause
                Bico
            end
            fprintf('\n\n Press any key to continure to the next example.')
            pause
        else
            if hit==2
                fprintf(' Correct!')

            else
                fprintf(' Wrong answer. The right answer was ''Rooms A and C are the same''.\nPress any key to
                 see that example again.')
                pause
                Bico
            end
        end
    end
    fprintf('\n\n End of Introduction. Press any key to continue.')
    pause
end
dir1=dir([curdir '1\']);
dir2=dir([curdir '2\']);
dir3=dir([curdir '3\']);
dir4=dir([curdir '4\']);
name_order=zeros(1,4);
clc
subject=input('Please enter your name: ','s');
fid=fopen(strcat('Experiment_B_score_',subject,'.txt'),'w'); %Creates a file for the subject
fprintf(fid,'%s \t\t %s \r\n',subject, date);
for r= 1:parts
    fprintf(fid,'PART %d \r\n',r);
    fprintf(fid,'%s \t %s \t %s \t Correct (=1) \r\n\r\n',['File A' blanks(26)],['File B' blanks(34)],
     ['File C' blanks(26)]);
    Combination_order=latin(12);
    for t = 1:12
        hit=0;
        name_order=randperm(4);
        names(name_order(1))=cellstr([dir1(Combination_order(t)+2).name]);
        names(name_order(2))=cellstr([dir2(Combination_order(t)+2).name]);
        names(name_order(3))=cellstr([dir3(Combination_order(t)+2).name]);
        names(name_order(4))=cellstr([dir4(Combination_order(t)+2).name]);
        if strncmp(names(1),names(2),17)
            A(Combination_order(t))=1;
        end
        if strncmp(names(1),names(3),17)
            A(Combination_order(t))=2;
        end
        if strncmp(names(1),names(4),17)
            A(Combination_order(t))=3;
        end
        names(name_order(1))=cellstr([curdir '1\' char(names(name_order(1)))]);
        names(name_order(2))=cellstr([curdir '2\' char(names(name_order(2)))]);
        names(name_order(3))=cellstr([curdir '3\' char(names(name_order(3)))]);
        names(name_order(4))=cellstr([curdir '4\' char(names(name_order(4)))]);
        clc
        sample_nameA=char(names(1));
        sample_nameB=char(names(2));
        sample_nameC=char(names(3));
        sample_nameD=char(names(4));
        SizeA=WAVREAD(sample_nameA,'size');
```

```
        SizeB=WAVREAD(sample_nameB,'size');
        SizeC=WAVREAD(sample_nameC,'size');
        SizeD=WAVREAD(sample_nameD,'size');
        [YA,FS,NBITS]=WAVREAD(sample_nameA,SizeA(1));
        [YB,FS,NBITS]=WAVREAD(sample_nameB,SizeB(1));
        [YC,FS,NBITS]=WAVREAD(sample_nameC,SizeC(1));
        [YD,FS,NBITS]=WAVREAD(sample_nameD,SizeD(1));
        pA = SND_MULTI([1 0 44100 16],ones(0,100));
        pB = SND_MULTI([1 0 44100 16],ones(0,100));
        pC = SND_MULTI([1 0 44100 16],ones(0,100));
        pD = SND_MULTI([1 0 44100 16],ones(0,100));
        Bico
        if hit==A(Combination_order(t))
            correct(Combination_order(t))=1;
        else
            correct(Combination_order(t))=0;
        end
    end
    for t=1:12
        fprintf(fid,'%s \t %s \t %s \t %s \t %d \r\n',[dir1(t+2).name blanks(46-length(dir1(t+2).name))],
         [char(dir2(t+2).name) blanks(46-length(char(dir2(t+2).name)))],[char(dir3(t+2).name)
          blanks(46-length(char(dir3(t+2).name)))],[char(dir4(t+2).name)
          blanks(46-length(char(dir4(t+2).name)))],correct(t));
    end
    if r ~= parts
        clc
        fprintf('End of part %d. You have answered %d out of 8 questions correctly.\n\nPlease take a
         few minutes break and when ready press any key to continue.\n',r,sum(correct(:)))
        dlmwrite(strcat('Experiment_B_Correct_part1_',subject,'.txt'),correct,'\t');
        pause
    end
end
clc
fprintf('You have answered %d out of 8 questions correctly.\n\nThank you!\n',sum(correct(:)))
fclose(fid);
%export matrices
dlmwrite(strcat('Experiment_B_Correct_part2_',subject,'.txt'),correct,'\t');
```

Bibliography

[1] Rasmussen, B., Rindel, J.H. and Henriksen, H., "Design and Measurement of Short Reverberation Times at Low Frequencies in Talks Studios", J. Audio Eng. Soc., Vol. 39, No. 1/2, 1991 Jan/Feb, 47-57.

[2] Børja, S. E., "How to Fool the Ear and Make Bad Recordings", J. Audio Eng. Soc., Vol. 25, No. 7/8, 1991 Jul/Aug, 482-489.

[3] Newell, P. R. and Holland, K. R., "A Proposal for a More Perceptually Uniform Control Room for Stereophonic Music Recording Studios", 103rd Convention of the Audio Eng. Soc., (Preprint no. 4580), 1997.

[4] Walker, R., "Early Reflections in Studio Control Rooms: The Results from the First Controlled Image Design Installations", 96th Convention of the Audio Eng. Soc., Amsterdam, Paper P12-6 (Preprint no. 3855), 1994.

[5] Walker, R., "A New Approach to the Design of Control Room Acoustics for Stereophony", 94th Convention of the Audio Eng. Soc., (Preprint no. 3543), 1993-03.

[6] Rec. ITU-R BS.1116-1, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including multichannel Sound Systems", 1994-1997.

[7] EBU Tech. 3276, "Listening conditions for the assessment of sound program material: monophonic and two-channel stereophonic" - 2nd Edition, May 1998.

[8] EBU Tech. 3276-E, "Listening conditions for the assessment of sound program material, Supplement 1, Multi-channel sound " - Pre-Press version, Feb. 1999.

[9] Spikofski, G., "Assessment of Sound Field Parameter Differences in Studio Listening Conditions", EBU Technical Review, Sep. 2000.

[10] Rindel, Jens Holger, "Introduction to Room Acoustics", Note 0114, Ørested, DTU, 2001.

[11] Jacobsen, Finn, "The Sound Field in a Reverberation Room", Note 2215, Ørested, DTU, 2003.

[12] Kuttruff, Heinrich, "Room Acoustics", Taylor & Francis; 4th edition (October 2000).

[13] Walker, R., "Room Modes and Low Frequency Responses in Small Enclosures", 100th Convention of the Audio Eng. Soc.,(Preprint no. 4194), 1996-05.

[14] Kuttruff, H., "Sound Fields in Small Rooms", AES 15th International Conference, Copenhagen, 1998, p 11-15.

[15] Vorländer, M., "Objective Characterization of Sound Fields in Small Rooms", AES 15th International Conference, Copenhagen, 1998, p 16-23.

[16] Rindel, J. H., "A New Method to Measure Coloration in Rooms Using Cepstrum Analysis", ICA 14, Beijing 1992, F3-4.

[17] Meynial, X. and Vuichard, O., "Objective Measure of Sound Colouration in Rooms", Acta Acustica, Vol. 85, 1999, 101-107.

[18] Watkins, Anthony J., "Coloration and Speech Perception", Acoustics Bulletin, 17-21, Nov/Dec, 1999

[19] Olive, S. E. and Toole F. E., "The Modification of Timbre by Resonances: Perception and Measurement", J. Audio Eng. Soc., Vol. 36, No. 3, 1988 March, 122-141.

[20] Olive, S. E. and Toole F. E., "The Detection of Reflection in Typical Rooms", J. Audio Eng. Soc., Vol. 37, No. 7/8, 1989 July/August, 539-552.

[21] Bech, S., "Timbral aspects of reproduced sound in small rooms, I", J. Acoust. Soc. Am., 97, 1995, 1717 - 1726.

[22] Bech, S., "Timbral aspects of reproduced sound in small rooms, II", J. Acoust. Soc. Am., 99, 1996, 3539 - 3549.

[23] Bech, S., "Spatial aspects of reproduced sound in small rooms", J. Acoust. Soc. Am., 103, 1998, 434 - 445.

[24] Bech, S., "Perception of Timbre of Reproduced Sound in small rooms: The Influence of the room and the loudspeaker position", J. Audio Eng. Soc., 42, 1994, 999 - 1007.

[25] Beranek, Leo L., "Music, Acoustics & Architecture", John Wiley and Sons Inc., 1962.

[26] Niaounakis, T.I. and Davies W. J., "Perception of Reverberation Time in Small Listening Rooms", J. Audio Eng. Soc., Vol. 50, No. 5, 2002 May, 343-350..

[27] Müller, S. and Massarani, P., "Transfer-Function Measurement with Sweeps", J. Audio Eng. Soc., Vol. 49, No. 6, 2001 June, 443-471.

[28] ISO/CD 3382-2, International Standard, "Acoustics - Measurement of the Reverberation Time - Part 2: Ordinary Rooms", 2003.

[29] ISO/R 1996-1971 (E), International Standard, "Acoustics – Description and measurement of environmental noise – Part 1: Basic quantities and procedures", Appendix Y.

[30] Blazier, Warren E. Jr., "Part 2: Noise Control Criteria for Heating, Ventilating and Air-Conditioning Systems", 7.49-7.62, from "Noise Control in Buildings - A Guide for Architects and Engineers" edited by Cyril M. Harris, McGraw-Hill Inc. 1994.

[31] Hammershøi, D., Møller, H. and Sørensen M. F., "Head-Related Transfer Functions: Measurements on 40 Human Subjects", xxth Convention of the Audio Eng. Soc., (Preprint no. 3289), 1992.

[32] Hammershøi, D., Møller, H. and Sørensen M. F., "Transfer Characteristics of Headphones: Measurements on 40 Human Subjects", xxth Convention of the Audio Eng. Soc., (Preprint no. 3290), 1992.

[33] Møller, H., Hammershøi, D., Jensen C. B. and Sørensen, M. F., "Transfer Characteristics of Headphones Measured on Human Ears", J. Audio Eng. Soc., Vol. 43, No. 3, 1995 April, 203-216.

[34] Pralong, D. and Carlile, S., "The role of individualized headphone calibration for the generation of high fidelity virtual auditory space", J. Acoust. Soc. Am., Vol. 100 (6), Dec, 1996, 3785-3793.

[35] Møller, H., Hammershøi, D., Jensen C. B. and Sørensen, M. F., "Evaluation of Artificial Heads in Listening Tests", J. Audio Eng. Soc., Vol. 7, No. 3, 1999 March, 83-99.

[36] Møller, H., Minnaar, P. and Christensen, F., "Localization with Binaural Recordings from Artificial and Human Heads", J. Audio Eng. Soc., Vol. 49, No. 5, 2001 May, 323-336.

[37] Møller, H., Hammershøi, D., Jensen C. B. and Sørensen, M. F., "Binaural Technique: Do We Need Individual Recordings?", J. Audio Eng. Soc., Vol. 44, No. 6, 1996 June, 451-469.

[38] Rumsey, F, "Spatial Audio", Focal Press, 2001, chapters 3 and 5.

[39] Poulsen, Torben, "Acoustic Communication - Hearing and Speech", DTU, 2003.

[40] Poulsen, Torben, "Psychoacoustic Measuring Methods", DTU, 2002.

[41] Guilford, J.P., "Psychometric Methods", McGraw-Hill Book Company, 1954.

[42] Gaver, William W., "What in the World Do We Hear?: Am Ecological Approach to Auditory Event Perception", Ecological Psychology, 5(1), 1-29, 1993.

[43] Bartz, Albert E., "Basic Statistical Concepts in Education and the Behavioral Sciences", Burgess Publishing Company, 1976.

[44] Wuensch, Karl L., "Statistical Lessons", http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm.

[45] Dorak, M.Tevfik, "Common Concepts in Statistics", http://dorakmt.tripod.com/mtd/glosstat.html.

[46] Whitehead, John, "An Introduction to Logistic Regression", http://personal.ecu.edu/whiteheadj/data/logit/.

[47] "Documentation for MathWorks Products", Release 14, 1994-2004 The MathWorks, Inc.

[48] Labgruppen Lab 300 Specification sheet:
http://www.labgruppen.se/media/LAB300_SPEC2.pdf

[49] ISO 3382:1997(E), International Standard, "Acoustics - Measurement of the reverberation time of rooms with reference to other acoustical parameters".

[50] Asgaard, Jesper, "Regulering af efterklangstid ved lave frekvenser", The Acoustic Laboratory, DTU, 1995.

[51] Gade, A. C., "Investigations of Musicians' Room Acoustic Conditions in Concert Halls. Part I: Methods and Laboratory Experiments", Acustica, Vol. 69, 1989, 193-203.

[52] Gade, A. C., "Investigations of Musicians' Room Acoustic Conditions in Concert Halls. Part II: Field Experiments and Synthesis of Results", Acustica, Vol. 69, 1989, 249-262.

[53] Hameed, S., Pakarinen, J., Valde, K. and Pulkki, V., "Psychoacoustic Cues in Room Size Perception", AES 116th convention Berlin, paper 6084, 2004.

[54] Sandvad, J., "Auditory Perception of Reverberant Surroundings", J. Acoust, Soc. Am. Vol. 105, No.2, Pt. 2, Feb 1999, 1193.