Antti Kuusinen

Perception of Concert Hall Acoustics - Selection and Behaviour of Assessors in a Descriptive Analysis Experiment

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Otaniemi 9.9.2011

Thesis supervisor:

Adjunct professor Tapio Lokki

Thesis instructor:

M.Sc Jukka Pätynen

Aalto University
School of Electrical
Engineering

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Antti Kuusinen

Työn nimi: Konserttisalin akustiikan havainnointi - Koehenkilöiden valinta ja käyttäytyminen kuvailevassa analyysikokeessa

Päivämäärä: 9.9.2011 Kieli: Englanti Sivumäärä:7+86

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Akustiikka Koodi: S-89

Valvoja: Dosentti Tapio Lokki

Ohjaaja: DI Jukka Pätynen

Tämä työ käsittelee konserttisalien akustiikan havainnointia. Työssä esitetään kolmen suomalaisen konserttisalin akustiikan tutkimuksen toteutus, joka koostuu koehenkilöiden valintaprosessista sekä menetelmästä, jossa jokainen koehenkilö kehittää oman kuvailevan sanaston näytejoukon arvioimiseksi. Lisäksi, työn kirjallisuuskatsauksessa esitetään tutkimusalan tämän hetkinen tila sekä käsitellään havaintotestien ja erityisesti kuvailevan analyysitekniikan metodiikkaa ja siihen liittyvää tulosten analysointia.
Tulosten osalta työssä keskitytään yksittäisten koehenkilöiden käyttäytymiseen, jota analysoidaan pääkomponenttianalyysin avulla tuotetuilla havaintoprofiileilla sekä *beta*-kertoimella, jolla voidaan arvioida profiilien monimutkaisuutta. Tulokset osoittavat merkkejä koehenkilöjen eroavaisuuksista kyvyssä jakaa havaintokokonaisuus sen tärkeimpiin elementteihin ja tuottaa kuvailevia attribuutteja, jotka antavat moniuloitteista informaatiota näytejoukosta. Kokonaisuudessaan tutkimus osoittaa, että kuvaileva analyysimenetelmä sopii hyvin myös konserttisaliakustiikan tutkimukseen ja sillä voidaan saada tarkkoja tuloksia.

Avainsanat: Akustiikka, konserttisalit, havainnointi, deskriptiivinen analyysi, pääkomponenttianalyysi, koehekilöiden valinta, metodiikka, havaintoprofiilit

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Antti Kuusinen

Title: Perception of Concert Hall Acoustics - Selection and Behaviour of Assessors in a Descriptive Analysis Experiment

Date: 9.9.2011 Language: English Number of pages:7+86

Department of Signal Processing and Acoustics

Professorship: Acoustics Code: S-89

Supervisor: Adjunct professor Tapio Lokki

Instructor: M.Sc Jukka Pätynen

This thesis focuses on the subjective perception of concert hall acoustics with an investigation of three Finnish concert halls by a novel descriptive analysis method referred to as individual vocabulary profiling. The literature study presents the current state of concert hall acoustics research as well as the methodological background of sensory analysis. The practical part consists of the screening of assessors and the implementation of the procedure. Data analysis involves the investigation of the individual behaviour of assessors with principal component analysis as well as the *beta*-coefficient, which is used to evaluate the complexity of sensory profiles. The results indicate clear differences between assessors in their ability to break the perception of acoustics into its constituting elements and to produce attributes which discriminate the stimuli in different perceptual aspects. The most prominent perceptual characteristics discovered in this study are distance, loudness, reverberance, definition and clarity. These findings are well in line with the previous studies of concert hall acoustics. Moreover, this study shows that this methodology can be well adapted to the field of concert hall acoustics and it yields accurate results.

Keywords: Concert halls, Acoustics, Psychoacoustics, Perception, Descriptive analysis, Sensory profiling, Data analysis

Acknowledgements

I would like to thank my family, friends and colleges for support and good advice during writing this thesis. Particularly, I would like to thank Tapio Lokki for the opportunity to participate and work on this research and for giving great insight on the subject. Also, many thanks go to the co-workers at the lab for discussions and ideas.

I need to say that this work has been written in many extraordinary places, from the beautiful beaches of Portugal to the wild banks of the River Teno up north in Lapland. It has been a long process, and I'm thankful for all the people who I have had the pleasure to meet and spend time with during this time.

Otaniemi 9.9.2011

Antti Kuusinen

Contents

Symbols and abbreviations

Sound and acoustics

RT	Reverberation time
EDT	Early decay time
ITDG	Initial-time-delay gap
D_{50}	Definition
C_{80}	Clarity
G	Loudness/Strength factor/Sound level
IACC	Interaural cross-correlation coefficient
SI	Spatial impression
ASW	Auditory source width
LEV	Envelopment
BQI	Binaural quality index
LF	Lateral fraction
DirAC	Directional audio coding

Methodology

JND	Just-noticeable-difference
HTL	Hearing threshold level
DA	Descriptive analysis
CVP	Consensus vocabulary profiling
IVP	Individual vocabulary profiling
MDS	Multidimensional scaling
RGT	Repertory Grid Technique
FCP	Free-Choice Profiling
FP	The Flash Profile
PCA	Principal component analysis
(H)MFA	(Hierarchical) Multiple factor analysis
GPA	Generalized procrustes analysis
ANOVA	Analysis of variance
GUI	Graphical user interface

1 Introduction

The science of architectural acoustics was founded by W. C. Sabine in the early 1900s and concert hall acoustics holds a special position in this vast field. Concert hall is the cradle of live acoustic music, where music can be fully enjoyed and appreciated by both the listeners and the performers. Sabine was the first to systemically investigate the properties of a hall, what resulted in defining the most important acoustical parameter up to date: reverberation time, RT. This discovery launched the science of architectural acoustics, which is now a multidisciplinary field spanning from basic physics to the complex systems of human auditory perception and psychology.

Since the times of Sabine, great insight has been gained on the particular properties of a hall required to attain perceptually satisfactory acoustics. It is well known, that acoustics should support the music, so that, it is engaging and immersing, different sounds should be clearly perceivable and the loud sounds should make their impact while the soft parts should also keep their intricacy. Also, the hall should neither be too "dry", or too "live", having a reverberation time around 2 seconds.

Many studies of subjective perception of concert hall acoustics have been conducted with questionnaires and in-situ listenings of real concerts while others have been carried through in laboratory circumstances using recordings or acoustical simulations. The in-situ listening offers the advantage of natural conditions, but there are many variables, which can not be controlled and making direct comparisons reliably is not possible. The listening tests in laboratory, in turn, offer controllable conditions and the possibility to perform tests of different kinds, but various issues of authenticity and artifacts are always present. In order to minimise this kind of tradeoff, the development of novel techniques to recording, simulation and creation of stimuli and more elaborate experimental designs are still needed.

Descriptive analysis methods commonly applied in the field of sensory science offer an interesting alternative to the more traditional listening tests. These methods are often used in product evaluation in order to obtain detailed descriptions of the stimuli or to investigate consumer preferences. Often these tests are performed by using a panel of assessors, but this approach has been found to require a lot of resources, time and commitment from both the experimenter and the assessors. Thus, there has been an emergence of the so-called individual vocabulary methods, where each assessor develop an own attribute list for the evaluation of stimuli. In the audio field, there are already promising results of descriptive analysis experiments, but in the particular field of acoustics, these methods are still quite unknown.

This thesis presents a part of an ongoing research of the Virtual Acoustics research group of Department of Multimedia of the Aalto University School of Science. In overall, this research includes the development of a controllable virtual orchestra consisting of 34 loudspeakers with recording and signal processing techniques as well as investigation of novel listening test methods, which can be paired with the technical advancements and innovations. This thesis concentrates on the assessor selection procedure and the implementation of the descriptive analysis experiment: the individual vocabulary profiling of three Finnish concert halls. Moreover, the focus here is on the individual behaviour of assessors in terms of individual sensory

profiles. The overall results of this study are presented elsewhere.

This thesis is organized as follows. Chapter 2 presents a literature study of concert hall acoustics including the key concepts of concert hall acoustics, architectural aspects, as well as the main studies of subjective perception of concert hall acoustics with a historical perspective. Chapter 3 discusses the methodological background of this work with a particular focus on the descriptive analysis methods. Also, various data analysis techniques are covered at the end of the chapter. Chapter 4 presents the implementation of the individual vocabulary profiling of three Finnish concert halls. This presentation consists of the screening of assessors, the implementation of the IVP procedure and investigation of the individual behaviour of assessors in terms of the individual sensory profiles. Finally, Chapter 5 gives a brief summary of the whole thesis.

2 The development and the current state of concert hall acoustics research

Concert hall acoustics have been studied for over a century. Currently, this multidisciplinary branch of acoustics research is a vivid and interesting combination of studies focusing on a diverse but interleaving aspects of acoustics. Concert hall acoustics research can be roughly categorized into studies focusing on the physically measurable quantities of sound, the effects of architectural solutions and acoustical treatment, the simulation, prediction and estimation techniques and studies which concentrate more on the psychoacoustical domain, particularly on listener's experience and auditory perception in concert halls.

Wallace Clement Sabine (1868-1919) can be regarded as the father of the science of architectural acoustics. As a young Harvard assistant professor of physics, his work on acoustics started by investigating the problems in a newly built university lecture hall and after several years of research in 1898, he derived the equation:

$$\mathrm{RT} = 0.16\frac{V}{A\alpha}, \tag{1}$$

where RT is the time it takes for a sound to attenuate to inaudibility, V is the volume of the space in cubic meters and A is the total surface area and α is the average absorption of the surfaces. Sabine's equation connected the reverberation time RT to two key variables, i.e., volume and absorption, and it is presented here because it still remains a foundation of architectural acoustics today. Later in 1900, Sabine helped to design the renowned Symphony Hall in Boston. [15]

Since the days of Sabine, advancements in measurement techniques have provided means to collect evidence on more complex acoustical phenomena than the reverberation time. Such measures include early decay time (EDT), initial time-delay gap (ITDG), clarity (C_{80}), loudness (G) and the interaural cross-correlation coefficient (IACC). These are discussed more in detail in the next section.

Knowledge gathered about human auditory perception has also helped to understand the psychoacoustical phenomena in concert halls. However, conducting subjective listening experiments with direct relation to concert hall acoustics that would have high scientific rigour has been proven to be often difficult. Thus, novel methods, like the one presented in this thesis, are still needed to specify the link between physical measures and the listening experience in concert halls.

The next paragraphs present an overview of the concert hall acoustics research up to date. To begin with, the key concepts in acoustics research are covered with the definitions of the most important physical measures of acoustical quality. Next, concert hall architecture is discussed briefly with a historical perspective. Then, the emphasis moves on to the studies investigating the listening experience and auditory perception and the connection between the objective measures and the subjective experience in a concert hall.

2.1 Key concepts of concert hall acoustics

Over the years, sound engineers and acousticians have developed and elaborated a range of physical measures for the description and comparison of the acoustical properties of concert halls. The multitude of these measures alone point to the complexity of the acoustical phenomena and multidimensionality of the listening experience in a concert hall. None of these measures could be used independently to fully characterize the acoustics of a hall, or the subjective perception. Many of these parameters have been originally developed in the very context of concert halls, but the same concepts can mostly be applied to estimating and measuring any enclosed space. In general, the objective parameters relate to sound propagation, sound field and reflection theories combined with the functions of the human auditory system and auditory perception.

Physical acoustical measures are not in the focus of this thesis. Still, they are inextricably related to the subject and provide a natural introduction the field of architectural acoustics. Thus, a list and descriptions of the main objective parameters are presented next. The mathematical equations are not presented here in this thesis, but they can be found in literature, e.g. [8]. Instead, the focus here is on the descriptions of subjective perception of these parameters in the context of concert halls. In the last section of this chapter, a more detailed discussion about this topic is provided. A more complete list and chronological order of all parameters developed over the years with the original references has been presented by Lacatis et al. [60].

Most of the acoustical parameters presented next, are tightly related to the concepts of direct, early and reverberant sound. In the domain of concert hall acoustics, direct sound is the sound that travels directly from an instrument on a stage to the listener seated in the audience. The term early sound includes the direct sound and all the reflections from all the boundaries, mainly walls and ceiling, reaching the listeners position in the first 80 milliseconds after the arrival of the direct sound. Finally, the reverberant sound encompasses all the reflections arriving to the listener's ears after 80 ms. [13] It is important to keep these basic concepts in mind while considering following acoustical measures:

- Reverberation time (RT):

  The first developed and the most extensively studied objective parameter is reverberation time, RT. It was formally presented for the first time by W. C. Sabine in the 1920s [89]. Reverberation time refers to the time period in which the sound attenuates to inaudibility (technically 60 decibels) after the source has stopped generating a sound. Reverberation is a product of a large number of echoes building up, bouncing between the surfaces of the hall and slowly decaying as the sound attenuates by the inverse-square law of the distance from the source and is absorbed by the surfaces and the air. Thus, reverberation time is highly dependent on the volume of the space as well as the surface materials and amount of acoustical treatment applied in the space as stated by Sabine. [13]

If the reverberation time is too short, a concert hall can be described as "dead" or "dry" and orchestral or symphonic music is not adequately supported by the hall. If the reverberation time is too long, the acoustics may be perceived as too "live" and the music as distant or lacking presence, clarity and strength. It is now well established that in concert halls with appraised acoustics, the RT is around 1.8 - 2.2 seconds depending on the purpose and shape of the hall. RT estimation and measurement techniques have been further elaborated, for example, by Schroeder [92], Ratman et al. [88] and Beranek [14].

- Early decay time (EDT):

  Early-decay time, originally presented by Jordan in 1970 [53], is closely related to reverberation time. It refers to the initial phase of the sound decay. To be exact, EDT is the time period in which the sound decays 10 dB after it is cut off, multiplied by a factor of 6. The reason for this multiplication is that the time it takes for the 10 decibel decay is roughly equivalent to the one sixth of the time for 60 dBs attenuation, previously described as reverberation time. Thus, multiplication by a factor of 6 allows a direct comparison between EDT and RT. However, in orchestral and symphonic music, successive notes are often played rapidly, especially by violinists, which means that only the early part of the sound decay process is audible. That said EDT is generally regarded as a better indicator of acoustical quality than RT. [13]

- Definition (D_{50}) and clarity (C_{80}):

  In orchestral music, several instruments play together and notes often follow each other at a high rate. Definition or clarity (C_{80}) refers to the degree to which a listener is able to distinguish individual instruments or individual notes in a musical performance. The former definition may be described as "vertical definition" and the latter as "horizontal definition" [13]. Vertical definition relates to the degree to which a listener can separate individual sounds that are played simultaneously while horizontal definition refers to how well a listener is able to separate sounds played in succession.

  In general, C_{80} is physically measured as the ratio of the energy in the early sound to that in the reverberant sound, expressed in decibels. Futhermore, the values of C_{80} are usually averaged over 500, 1000, and 2000 Hz octave bands and over several measurement positions. When C_{80} has a large positive value, the room is very dead or dry and early sound dominates, and music may be described as clear. When the room is very live, C_{80} normally has a large negative value and music is often perceived as "muddy". Note, that the C_{80} is highly negatively correlated with RT and EDT. [13]

  According to Lacatis et al. [60] the term "definition" was originally presented by Thiele in the 1950s and "clarity" by Alim in the 1970s. However, currently acousticians treat these terms mostly as synonyms.

- Strength factor G and loudness:

The term 'loudness' is one of the most important parameters and it refers to the strength of the sound in a concert hall. As the dynamic range in classical orchestral music can be very large, it is important that the concert hall supports both the quiet and subtle sounds in pianissimo parts and the loud sound levels at the fortissimos. Loudness of a hall is objectively described with the strength factor G in decibels, which was originally presented by Lehmann in 1976 [66]. It is a measure of the sound-pressure level at a point in a hall, with an omnidirectional source on stage, minus the sound pressure level of the same sound source measured at a distance of 10 m in free field.

Overall G is calculated by averaging the measurements over all octave frequency bands, but G can also be associated with a particular frequency range. The two most often encountered G parameters are G_{mid}, which is G average of the measurements in the 500 and 1000 Hz octave frequency bands, and G_{low} or G_{125}, which refer to the absolute strength of the sound in the lower frequency bands or at 125 Hz respectively. G_{low} and/or G_{125} are often used to describe the strength of bass or the perceived warmth in a concert hall [13].

- Initial-Time-Delay Gap, ITDG:

  Initial-time-delay gap was first introduced by Davis [30] in the context of control room design. Later Beranek [13] initiated its use as a measure of acoustical intimacy in concert halls. Intimacy may be defined as the subjective impression of the "closeness" of music or performance even when the real physical distance is large. Physically, ITDG refers to how soon after the direct sound the first reflection arrives to the listener's ears, i.e., to the time difference between the direct sound and the first reflection. If ITDG is short, the concert hall may be described as "intimate" or that the hall has "presence". In the best-liked halls, ITDG measured at the center of the main floor is usually at or below 25 ms. [13]

- Measures of auditory spatial impression and spaciousness:

  The overwhelming experience of being immersed in music while listening to a performance in a concert hall has been one of the main topics of the studies of concert hall acoustics. Barron [9] described this phenomenon with a term spatial impression (SI) with a quote from Marshall: "The sensation of spatial impression corresponds to the difference between feeling 'inside' the music and looking 'at' it, as through a window." Since, as discussed by Griesinger [41], several authors have described the auditory spatial impression with at least two forms of spaciousness: the auditory source width (ASW, also apparent source width), which refers to the perception of the width of the sound source on stage and listener envelopment (LEV) which refers to the surround effect of the reverberant sound. It has been observed that the early sound arriving before 80 ms after the direct sound contributes more to ASW, whereas later reflections are more associated with the listener envelopment.

  While different researchers and experts may be using different terminology for describing spaciousness, the importance of the lateral reflections in generating

this experience has been stated by all. Thus, it is primally the measurements of the lateral reflections or lateral energy which are applied to objectively describe the experience of the spatial impression and spaciousness. Such measures include the interaural cross-correlation coefficient (IACC), binaural quality index (BQI) and lateral fraction (LF), which are briefly discussed in the following.

The interaural cross-correlation coefficient is a measure of the similarity of the sound between the two ears. When the sound arrives directly from in front of the listener, the signals are exactly the same in both ears and IACC is equal to unity. When the sound signals are totally different between the ears, e.g., sound arrives from one side only, IACC value is close to zero. Thus, IACC is particularly affected by the lateral reflections bouncing of the side walls of the hall. IACC was first proposed by Schroeder et al. [93] in 1974.

The perception of the width of a source on stage, i.e., ASW is also greatly affected by the lateral reflections from side walls. IACC and ASW are inversely correlated, so that, when perception of source width is great, the signals between the ears are very dissimilar. ASW and, more generally, the subjective perception of spaciousness has been also associated with the acoustic quality of the hall. To enable positive correlation to the perceived acoustical quality, Beranek has proposed a quantity called Binaural Quality Index (BQI), defined as (1-$IACC_{E3}$) where E designates the early sound and "3" indicates the average of the $IACC_E$ values in the 500, 1000, 2000 Hz octave bands, to be used instead of IACC. [13]

Lateral fraction (LF) as a measure of the strength of lateral reflections was developed by Barron in 1971 [8]. LF equals to the ratio of the energy in lateral reflections to the total energy arriving at the listener position in a hall. LF is associated with the broadening of a sound source on stage beyond its visual width, and it is thus positively correlated with ASW) [13]

2.2 Concert hall architecture

Before the times of Sabine and the beginnings of the science of architectural acoustics, an acoustically successful concert hall was often simply a product of good luck. Today, the architectural features of concert halls have been extensively studied and the basic criteria for establishing desired acoustical quality are well known. In his book *"Concert halls and opera houses"* [13] Leo L. Beranek gives a full description of one hundred concert halls and opera houses around the world as well as the subjective rank-ordering of acoustical quality of 58 halls. At sight, it can be clearly noted that there are at least as many concert hall designs as there are architects designing the halls.

Architects are often very keen in keeping their designs original, so, each new concert hall is usually a unique construction. Moreover, the final design is always a compromise between architectural features, such as, the shape of the hall and the number of seats, and acoustical qualities, such as, reverberation time, clarity and loudness. Furthermore, the acoustics of a concert hall should be designed for a

particular musical style as the performance is greatly affected by the acoustics of the hall. Traditionally, concert halls have been designed for orchestral and symphonic music, but recently, there has been a growing trend in designing multi-purpose halls, which could accommodate not only the symphonic orchestras of various musical styles but also smaller chamber ensembles or even bands employing public address (PA) systems. One reason for this progression may well be seen in the number of people attending to popular music performance venues compared with the audience of classical concerts as reported by Adelman-Larsen et al. [1].

Beranek's rank-ordering of concert halls has been performed by analysing and interpreting the interviews and questionnaire surveys of the conductors, music critics and enthusiastics of concert music in conjunction with the author's own opinions. As it is also noted by Beranek himself, the people who participated in this survey were most often familiar with only a small fraction of the concert halls in question, and thus, this survey does not fully meet the requirements of a scientific work. However, it does establish an excellent overview on concert halls by combining their architectural solutions with the general opinions of their acoustical performance.

20 topmost ranked halls in Beranek's list were all built in the first decade of 20th century or before, the three at the top being Grosser Musikvereinssaal in Vienna (1870), Symphony Hall in Boston (1900) and Teatro Colón in Buenos Aires (1908). Interestingly, while there has been a giant leap in the construction and material techniques between 1900 and today, acoustical quality seems not to have improved. One could well argue that the technical restrictions in the building architecture over century ago may have ensured the acoustical quality, whereas the technical advancements have not only made possible to built larger and more complex performance spaces but also may have deteriorated the acoustical quality at the same time.

Two-thirds of the 15 highest ranked halls, are "shoebox" shaped halls (e.g., Musikvereissaal and Boston Symphony Hall). A shoebox, as it name indicates, is typically a symmetrical rectangular hall with a high ceiling and balconies on the sides and at the back of the hall. This simple architectural design is often acoustically a safe solution as the parallel side walls assure the early lateral reflections to the main audience area, which, as stated before, are essential for broadening the apparent source width and increasing the feeling of envelopment and spaciousness. However, one apparent drawback of this design is that musicians on stage may easily feel isolated from the audience.

In "surround" halls, audience is seated around the orchestra, usually in "trays" or terraces, enhancing the connection between the performers and the listeners. Acoustical quality in these halls is however more difficult to achieve as lateral reflections are not provided by the side walls. Many solutions include hanging reflectors from the ceiling above the orchestra to provide an adequate amount of early reflections and dividing seating areas with "walls" to improve the presence lateral reflections. Other non-rectangular designs are the "vineyard" shape, in which seats are positioned on sloping sections more or less around the orchestra, and the "fan" shape, which was a quite popular design in the late 20th century but later found to be acoustically problematic, as there is very little support from lateral reflections. Currently, also various multipurpose halls are being build, where the stage and the

audience area can be modified according to the performance.

2.3 Previous research on subjective perception of concert hall acoustics

This section is dedicated to the main studies, in which have been conducted subjective assessments of concert hall acoustics. In most of the studies presented here, investigations have also included evaluations of the objective parameters with attemps to correlate the physical measurements with the results from subjective experiments. All studies that have been included here, have concentrated particularly on concert hall or auditorium acoustics in a comprehensive way, whereas there also exist an excessive amount of work that have a more indirect relation to concert hall acoustics and studies that have focused on some particular aspect of the subjective auditory experience.

There are a few in-depth reviews of studies of subjective assessment of auditorium acoustics. The most cited might be the article *"Concert hall acoustics - 1992"* [15] by Beranek, which serves still as a quite thorough description of the state of art of the concert hall acoustics research. In addition, the main references for this section have been the theses of Catherine Lavandier [61] and Eckhard Kahle [54], which are unfortunately written in French and, thus, not so popularly cited in this field. An overview of the main studies is illustrated in Table 1.

2.3.1 Sabine (1900-)

As described already at the beginning of this chapter, W. C. Sabine was the founder of the science of architectural acoustics. He is mostly known for the famous mathematical equation for the reverberation time, which was the first quantitative description of auditorium acoustics. However, Sabine was very conscious of the multidimensionality of acoustical phenomena and that reverberation time RT was not alone sufficient to describe the total acoustical quality of a hall. In his article *"Reverberation"* [89] he writes:

> "In order that hearing may be good in any auditorium, it is necessary that the sound should be sufficiently loud; that the simultaneous components of a complex sound should maintain their proper relative intensities; and that the successive sounds in rapidly moving articulation, either of speech or music, should be clear and distinct, free from each other and from extraneous noises."

Subsequently, he calls these three factors that affect the auditory perception in a hall as:

1. Loudness,
2. Distortion of Complex Sounds: Interference and Resonance, and
3. Confusion: Reverberation, Echo and Extraneous Sounds.

Table 1: The subjective assessment of auditorium acoustics. [69]

Who	Year	Excitation	Recording / Reproduction	Method(s)	Analysis	Main Findings
Sabine	1900-	-	-	-	-	1. loudness, 2. interference and resonance, 3. reverberation and echos
Beranek	1960s-	live orchestra	in-situ listening	interviews	mapping with objective data	1. reverberance, 2. loudness, 3. spaciousness, 4. clarity, 5. intimacy, 6. warmth, 7. hearing on stage
Hawkes & Douglas	1970s	live orchestra	in-situ listening	16 semantic differential scales	factor analysis	1. reverberance, 2. balance and blend, 3. intimacy, 4. definition, 5. brilliance
Barron	1988	live orchestra	in-situ listening	questionnaire	correlations	G, EDT, LEF, Two preference groups: reverberance and intimacy
Kahle	1995	live orchestra	in-situ listening	questionnaire of 29 questions	PCA, correlations	8 descriptive factors
Berlin group	1970s	live orchestra	dummy-head / headphones	questionnaire	19 direct attribute scales	1. loudness (G), 2.clarity (Ts), 3. timbre (EDT ratio), Two preference groups: loud sound and clear sound
Göttingen group	1970s	anechoic music / 2 loudspeakers on real stage	dummy-head / 2 louspeakers	preference, paired comparison (equalized loudness)	factor analysis	negative correlation between distinctness and preferred consensus factor; RT, D50, IACC
Lavandier	1989	anechoic music / simulation	headphones	non-verbal dissimilarity method	INDSCAL	11-14 descriptive factors
Soulodre & Bradley	1995	anechoic music / measured BRIRs	2 loudspeakers	paired comparison, preference	correlation	1. clarity, 2. treble
Lokki et al.	2010	anechoic music / 34 loudspeakers on real stage	B-format / 16 loudspeakers	individual vocabulary development	AHC, LDA, (H)MFA, RDA	1.loudness and distance, 2. reverberance (2 groups), 3. definition, 4. apparent source width

These perceptual aspects, in one form or another, are also present in most of the subjective studies conducted since.

2.3.2 Beranek (1955-)

In 1962, Beranek published the book *"Music, Acoustics & Architecture"*, with a first evaluation of the acoustical quality of 58 halls. As already stated before, there are various issues of the scientific rigour in this study, but many of the ideas elaborated by Beranek have constituted the platform of a large number of studies conducted since. Unfortunately, the original publication has not been disposable for the writ-

ing of this thesis, thus the observations presented here are based on the theses of Lavandier [61] and Kahle [54] as well as on the article *"Concert hall acoustics - 1992"* [15] by Beranek himself.

The objective of Beranek's study was to compare the different halls in overall opposed to considering the acoustical quality in the different areas inside a hall. In addition, Beranek concentrated on developing a vocabulary for describing the musical acoustical quality in concert halls. To obtain such a language, he interviewed conductors, performers and music critics as well as gathered his own impressions of the concert halls. Based on these observations, which were made in fully occupied concert halls, with a full symphony orchestra performing at least one major work from classical or romantic period, he developed an original list of 18 subjective attributes for describing the acoustical quality of concert halls (see the original publication or [54] for full attribute list).

Since the original publication, Beranek's research continued and by combining the results from subsequent studies performed by several other researchers (see [15] for details), he elaborated the rank ordering of 58 concert halls. He devised a rating system of five independent subjective attributes for judgments made by listeners as well as two additional attributes related to stage and the performers' perceptions. The audience related attributes were "intimacy", "liveness or reverberance", "warmth", "loudness", and "diffusion". The two player- and stage-related attributes were "balance and blend" as well as "ensemble". Furthermore, in the summary of *"Concert hall acoustics - 1992"* he presents that there are seven basic and essential subjective attributes of concert hall acoustics which are to be considered in the concert hall design. These attributes are: (from [15])

1. Reverberance
2. Loudness
3. Spaciousness
4. Clarity
5. Intimacy
6. Warmth
7. Hearing on stage

Finally, in 1996 Beranek published the book *"Concert halls and opera houses - Music, acoustics and architecture"* (a revised edition published in 2004), which presents a full description of one hundred concert halls and serves as a comprehensive reference for all interested in concert halls and their architecture. He also further elaborates the language of musical acoustics with many more descriptive terms than those which had been included in his previous work.

Beranek's studies as well as the deficient acoustics of the new concert hall in New York (for the description of the sad early phases of this hall, see [15]) initiated

several other investigations of auditorium acoustics with the objective of overcoming the disadvantages of the judgments made by interviews and questionnaires. Kahle [54] have described these issues with the following aspects:

- Semantic issues. It is possible that different test subjects may describe the same acoustical phenomenon with different terms (or the same terms may well be used to describe different acoustical phenomena).
- The issues of long term memory of acoustics. The jugdments and comparisons of the halls that are based on the listening experiences obtained in months, or even, years intervals.
- Issues of other acoustical influences than those of the acoustical quality of the halls. The influences of the orchestra, the performance and the musical piece.
- Non-acoustical influences. Architectural aspects and the fame of the hall.

It is important to note that these issues have been also considered in performing the current study and the solutions found are discussed in detail later in this thesis.

2.3.3 Göttingen group (1965 - 1976)

One may argue that the principal objective in constructing a new concert hall is to make it acoustically pleasurable to the audience. Göttingen studies [93] performed by Siebrasse, Gottlob and Schroeder addressed this issue by conducting listening tests that concentrated on listeners' preferences. To overcome the disadvantages described in the previous section, they developed an experimental design, which already incorporated many of the basic ideas that have been elaborated also in the current study.

Göttingen studies were performed in three stages. Stereophonic anechoic recordings (part of Mozart's Jupiter Symphony) were played over two omnidirectional loudspeakers, 5 m apart, 3 m upstage of the stage front and 1 m above the stage in 25 unoccupied European concert halls. The sound was recorded binaurally with an artificial head (dummy-head) in one central main-floor position in 22 halls and in ten positions in three halls. For the subjective listening test, these second recordings in turn were reproduced in an anechoic chamber with two loudspeakers with a "cross-talk cancellation technique", which means that each ear heard only what it would if the sound was played over perfect earphones. Furthermore, the subjective listening levels of the recordings were equalised in the listening test.

The recordings were presented in pairs to the test subject, one after another and the task was to simply report the preference for one, or the other, or no preference. Issues in listeners performing the evaluation with adjectival categories were thus eliminated, but on the other hand, it disabled the possibility to know the basis on which the judgements had been made. Also, this evaluation method does not provide information on factors which do not influence the subjective preference but serve merely for the differentiation of the samples.

Furthermore, it is generally accepted that loudness affects the perception of spaciousness, particularly envelopment and reverberance. Consequently if spaciousness had been one of the main criteria of preference, equalizing the sound levels must have also influenced the judgments of preference. Finally, although many of the non-acoustical aspects (the influences of the orchestra, players, conductor etc.) may be controlled by using loudspeakers as sound sources, one can point out that only two loudspeakers are hardly sufficient to simulate a real orchestra on stage.

The subjective responses were analysed with multiple factor analysis, which revealed two to four principal dimensions. The first factor was interpreted as describing a consensus preference between test subjects and the others as describing differences between the personal opinions of the subjects. Finally, only the first factor was analysed in terms of correlations with objective criteria. Main results were that reverberation time RT, and the definition parameter D, were highly positively correlated with the global preference and IACC was negatively correlated. [15] [54] [61]

2.3.4 Berlin group (1968 - 1976)

The experimental design of a research group in Berlin, lead by Wilkens and Lehman, consisted of making binaural ("dummy-head") recordings of the Berlin Philharmonic Orchestra playing in six German concert halls. In each hall (unoccupied), the orchestra played the short extracts of three musical pieces (Mozart, Brahms and Bartok) and the recordings were made in several seating positions. The binaural reproduction for the listening tests was made through earphones, which enabled an instantaneous comparison of the samples. Nineteen sample pairs were presented to forty test subjects who were asked to compare and rate their impression of the acoustics of the halls on 19 category scales, each with 6 points. The scales and end-point labels are illustrated in Figure 1. [15]

Wilkens analysed the data with multiple factor analysis, which resulted in a three-dimensional factor space explaining 90 percent of the variances in these 19 variables. First dimension was interpreted as "strength" or "volume", the second as "distinctness" or "clarity" and third as "the timbre of total sound" or "spectral balance". In addition, by investigating the subjective preferences, two groups of test subjects of equal size with different preferences were found: one group preferred a "loud" sound and the other a "clear" sound. [15]

As Kahle [54] points out, these results substantiate the idea, that auditory perception is structured in perceptual factors or dimensions that are common for all, but the interindividual differences in perception manifest in preference judgements. Each listener prefers different "values" for different perceptual factors and different factors are given unequal significance by different listeners.

2.3.5 Dresden group (1966 - 1980)

In their studies, a research team in Dresden (Schmidt, Abdel Adim, Lehmann, Reichardt) used synthetic sound samples in order to have full control of the characteristics of the samples. They primally studied the "hall effect" and clarity. The

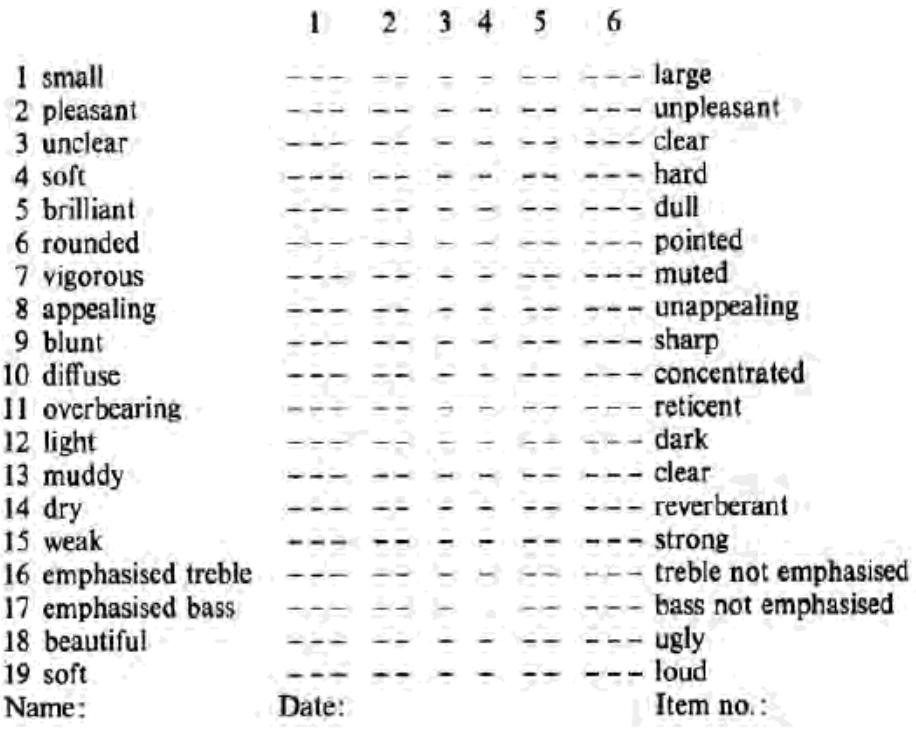


Figure 1: Scales and end-point labels used by Wilkens, translated into English. Picture taken from [54].

main results were that the hall effect could be divided into two different aspects: spaciousness and reverberance. They also found that separation into temporal (or horizontal) transparency and vertical transparency (how different sounds played at the same time can be distinguished) is futile as these two aspects are highly correlated. They also improved the correlations between the objective measures and subjective perception of hall/room effect (R), and clarity (C_{80}). [54]

2.3.6 Hawkes & Douglas (1971)

In the 1971, Hawkes and Douglas [46] presented a study in which they used a questionnaire in real concert situation. Four subjects assessed the acoustics of four concert halls with 16 semantic differential scales. By factor analysis, they reduced the results into four to six independent aspects depending on the case. These aspects were interpreted being reverberance, balance and blend, intimacy, definition, brilliance and proximity.

2.3.7 Barron (1988)

Barron [7] conducted a subjective evaluation of eleven concert halls in Britain with a group of expert listeners as assessors. The listeners attended real concerts and

were seated at least in two different positions in each concert - typically on one seat before an intermission and in another after the intermission. In each place, they were asked to evaluate the acoustics with a questionnaire of nine semantic differential scales. The attributes to evaluate were: clarity, reverberance, envelopment, intimacy, loudness, balance (treble, bass, the orchestra), background noise and the overall impression.

According to Kahle [54], two main aspects to be considered from this study are:

- Two groups of test subjects with different preferences were found: one group liked the best a great amount of reverberance, while the other preferred a good intimacy. These results correspond with the studies of the Berlin group.

- Contrary to the Berlin group studies, the distinction into two groups, could also be seen to some extend in the interpretation of the attributes. Particularly ambigous were the question of spatial envelopment; for the listeners who preferred great amount of reverberance, responses to envelopment were strongly correlated with the question of reverberance, whereas for the listeners who preferred greater intimacy, envelopment correlated strongly with the question of intimacy. Also, considering the objective parameters, listeners in the first group paid more attention to the spatial effect of the late reverberation, whereas in the second group, high correlation was found between envelopment and the objective sound level.

2.3.8 Lavandier (1989)

A series of listening tests and Catherine Lavandier's thesis [61] under the supervision of J.-P. Jullien was conducted in the IRCAM research group in France. The objective of these studies were to identify a set of perceptual factors and the corresponding objective criteria that could be considered as a basis of the multidimensional perceptual space. Lavandier concentrated on validating the common acoustical objective measures in terms of perception. The main objective criteria in focus were: reverberation time (RT), sound level (G), clarity (C_{80}), frequency bands of sound level and reverberation time as well as spatial distribution of early reflections. The sound samples in the listening tests were anechoic recordings (Bach and Bellini) manipulated with a set of delays and filters and a reverb unit in order to produce a controllable artificial hall effect.

Two different approaches were included in the 17 listening tests. Tests that focused on the temporal aspects were performed with headphones, whereas in the tests which investigated the spatial aspects, samples were reproduced over 11 loudspeakers in an anechoic chamber. In each test, one to three criteria were manipulated while the others were held as constant as possible. The task was to evaluate the dissimilarity between two samples in terms of a particular attribute. An average of 12 assessors and 8 different configurations were included in the listening tests.

The data were analysed primarily by Individual Differences Scaling (INDSCAL), which yielded a total of 14 perceptual factors. These could be further categorised

into 4 groups: the temporal factors (5), the effects of early reflections (3), the effects of RT and G (4) and the spatial effects (2). [61, 54]

2.3.9 Soulodre & Bradley (1995)

Soulodre and Bradley [21] also used anechoic recordings but they produced the hall effect by convolving the recordings with measured binaural room impulse responses from actual halls. With a double-blind paired comparison test method, ten test subjects were asked to evaluate the difference between the samples in terms of loudness, clarity, reverberance, bass, treble, envelopment, apparent source width, and overall preference. There were a total of 45 sample pairs produced by 10 different impulse responses measured in different concert halls.

By analysing the various correlations in the data, they investigated the relationships between the subjective ratings and the objective criteria. The main results were: (a) A-weighted sound level parameter ($G(A)$) is to be used instead of the traditional strength factor G, as it considerably improved the relationship with the subjective loudness ratings, (b) clarity is more tightly related to C_{80}, when relative sound level is combined in the parameter, (c) the perception of bass is dominated by the low frequency content of the early sound (50 ms), (d) the perception of treble is determined primarily by the high-frequency content of the late sound and (e) preference judgment were found to correlate with both clarity and treble. [21]

2.3.10 Kahle (1995)

After the tests in the laboratory (1986-1989) and the thesis of Lavandier (1989), the acoustic laboratory of IRCAM proceeded to a campaign of measurements and listening tests in nine European concert halls. During a period of two years, a group of about ten assessors evaluated the acoustics in real concert situations and in several seating positions in each hall with a structured questionnaire. Also, an excessive number of objective measurements were gathered from each hall. The goal of this campaign was, on the one hand, to collect research material and data for future studies and, on the other hand, to validate and finalize the researches administrated in the laboratory. This study and the main results are well described in the thesis of Eckhard Kahle [54], published in 1995.

A structured questionnaire was formed for the subjective assessment of the halls. A questionnaire was based on the perceptual factors obtained from the laboratory tests and on the first interpretation published in Lavandier's thesis. Finally, the questionnaire included 29 questions divided into five categories: (1) the perception of the acoustics (the sound level, dynamics, reverberation, envelopment etc.), (2) the general perception of sound sources (the subjective distance, the localisation and presence of instruments, contrast, definition and clarity etc.), (3)the perception of sound sources per instrument section (the subjective distance, the localisation and presence of instruments, definition and clarity etc.), (4) the spectral balance (the reverberance and level of low and high frequencies) and (5) the preferences and personal opinions (the general impressions, general balance and homogeneity, adaptation of the musical piece to the hall etc.).

One main contribution of Kahle's thesis is the analysis and investigation of different influences on the responses of the questionnaire. Kahle's approach consisted of isolating the influences of the listener, the musical piece, the place (i.e., the hall and the seating position), the interaction between musical piece and place and the residual noise and evaluating these influences separately. By this analytical approach, it was shown that it is possible to evaluate the acoustic quality in a detailed and reliable way in a real concert situation with a questionnaire although some concerns of the level of residual noise still remained. Furthermore, a global analysis of data resulted in the reduction of 29 questions to the 8 most relevant (fundamental) attributes which were described as: (1) sound level, (2) reverberance, (3) general balance, (4) contrast, (5) level of low frequencies, (6) level of high frequencies, (7) muddiness and (8) hardness (heurté).

The main results consisted of improving the correspondence between several perceptual factors and the objective measures. Concerning the current thesis, main results were:

- Sound level: Perception of sound level was separated into two different aspects: the presence of the source (i.e, the perception of early sound energy) and the presence of hall effect (i.e, perception of late sound energy).

- Reverberation: The subjective responses to reverberation varied significantly between the places inside a hall as well as between the halls. The traditional RT parameter explained some of the variations between halls although better correspondence was obtained with EDT. However, the correspondence with the variations inside a hall remained low for both parameters.

- Contrast: The perception of contrast was not related to only one single objective parameter but to three influencing factors: the sound level, particularly at high frequencies, the temporal fluctuations in early sound energy and the energy relations in the reverberant sound.

- General balance: This question was often related to the second dimension of multidimensional analyses and thus, could be considered being high on the hierarchy of attributes. However, any valid correspondence between this attribute and acoustical criteria were not found. It was concluded that several perceptual influences are incorporated in this question (e.g., spectral, spatial and instrumental balance). Thus, it may not be considered as a single perceptual factor.

- Muddiness: Muddiness was related to the lack of definition of one or several instruments. According to Kahle, this may have been due to unequalities of the frequency responses of the halls related to the early sound as well as to the late sound.

- Hardness: This aspect was related to punchy or hard sound, lacking of fluency. It was found that a low value of this aspect was preferred both in real concerts and in laboratory tests. According to Kahle, it could be related to a lack

diffusion in a concert hall and an objective criterium characterizing this effect would be needed.

- Subjective preference: It was found that preference judgments were influenced by all of 8 fundamental questions, thus, it was concluded that subjective preference is determined by the quality of several factors at the same time.

2.3.11 Other related studies

David Griesinger can be regarded as one of the main contributors in the field of acoustics. He has worked on various aspects of room acoustics but concerning the concert halls, his main contributions have been the several studies and papers on auditory spatial impression, apparent source width, localization, envelopment, reverberance and warmth in halls and performance spaces, see [38, 39, 40, 44, 41, 42, 43]. He has particularly elaborated acoustical parameters from the perspective of psychoacoustics, the starting point being the functionalities of human hearing and auditory system (e.g., localization). The home page of Griesinger [37] is a good information source on his work and also provides information and comments on his more recent studies. One interesting and relevant topic to this thesis is the elaboration of the attribute "engagement" in concert halls (see slides "The Relationship Between Audience Engagement and Our Ability to Perceive the Pitch, Timbre, Azimuth and Envelopment of Multiple Sources" at [37]. However, a formal paper on this topic is yet to come.

Ando's preliminary work in Göttingen and in Kobe, Japan has been extensively reviewed by Beranek [15]. The most relevant studies for the current thesis are the subjective listening tests, where listeners were exposed to sound fields containing the direct sound, reflected sound waves from various directions and at various sound levels, and subsequent various reverberation fields (e.g., [2, 3]). The objective of these studies were to find the orthogonal factors of subjective preference for sound fields. Results yielded four orthogonal physical factors described as: 1) the listening level, 2) the initial time delay gap, 3) the subsequent reverberation time and 4) the interaural cross correlation coefficient. These results have been elaborated and resulted in a theory of individual preferences of sound fields in a concert hall [6], which was later applied in a design process of a new concert hall [5]. One of the main aspects of this theory is its association with the cerebral hemispheres of the human brain. In addition to this work, Ando, in collaboration with others, has elaborated a model of auditory brain system [4], which incorporates the autocorrelation mechanisms, the interaural cross-correlation mechanism between the two auditory pathways, and the specialization of the human cerebral hemispheres to the temporal and spatial factors of the sound field. He has also described how the subjective attributes of concert hall acoustics can be extracted based on this model. A more detailed discussion about this topic is however outside the scope of this thesis.

Besides Ando's work, there is a substantial amount of acoustics research conducted in Japan since the beginning of the 1970s, also reviewed by Beranek [15]. Unfortunately, many of these papers are written in Japanese, particularly earlier works like the one by Kimura and Sekiguchi in 1976, who, according to Beranek,

recorded the sound from a non-directional loudspeaker on stage in 13 Japanese halls with a dummy head on two seats in each hall. Sound was reproduced through earphones and the task was to evaluate the loudness, the quantity and quality of reverberation, spatial impression, brilliance, definition, proximity and overall preference. It was found that, preference could be explained mainly by the width of the hall and the cubic volume. Other Japanese studies reviewed by Beranek include the work by Nagata, Toyota et al. Nagamoto and many others. One can conclude from Beranek's review, as well as from the sheer amount of more recent studies and publications by Japanese researchers, that Japan, has been and is one of main regions of acoustical research at the moment.

A multi-institutional study of acoustic quality of auditoriums in Europe and Japan was conducted in 1986-89 by four scientific groups from Japan and one from Germany [97]. Acoustical measurements were made in fifteen European auditoriums and five in Japan. Also, seven of the European concert halls were included in preference tests of acoustical quality. Anechoic music was played over an omnidirectional loudspeaker on the stage of each hall and the sound was recorded with dummy head at a listening position at 12 m from the source. Later, music was reproduced over two loudspeakers facing the listener on an angle in an anechoic chamber. 88 listeners participated to the listening test, in which they were asked to give preference for one of a pair of sound fields. Interestingly, there was no significant preference among the seven halls when the results were averaged over all subjects, but the subjects could be divided into several groups according to preference. Moreover, it was concluded that the preference scores alone are not sufficient for evaluation of acoustical quality of concert halls.

There is still also an extensive amount of work performed by several researchers that lies outside of the scope of this thesis.

2.3.12 What have we learned?

The previous sections provide detailed descriptions of the main studies where a subjective assessment of concert hall acoustics has been performed in one form or another. These studies have started off from one experimenter's, Sabine's clear-sighted remarks on the multidimensionality of perception of acoustics and continued with the questionnaires and interviews conducted by Beranek, Barron, Kahle and others as well as with studies applying simulations, anechoic recordings, various hall recording and sound reproduction techniques and various listening test methods.

In overall, these studies, on the one hand, highlight the multidimensionality of the perception of concert hall acoustics and describe a set of attributes, which may be used for the acoustic evaluation, and on the other hand, exemplify the several issues to be considered in the subjective assessment of concert hall acoustics. Some of these issues have already been previously glossed over, but a more thorough discussion about these aspects is provided here.

Being perhaps the simplest method of subjective evaluation, questionnaires (with interviews) have been used by several researchers over the years. Advantages in using questionnaires are many: they can be used in real concert situations, in fully

occupied concert halls where the experience of acoustics is in the most natural form, the application is simple, straightforward and fast, and the additional qualitative information, that may be obtained by informal or formal interviews, is very valuable for the research. However, there are also several important drawbacks in using questionnaires including: semantic issues in the interpretation of questions or attributes, the number and selection of the items, issues of comparison and auditory memory and various issues of controlling the variables, assessor's mood etc.

To overcome the drawbacks in using questionnaires, several researchers have developed various experimental designs, in which semantic issues can be eliminated and as many variables as possible can be controlled. Besides the obvious limitations of not being able to perform the assessment in real concert situation, there are many advantages and drawbacks depending on the particular experimental design.

Acoustic simulations applied to anechoic recordings have been used by Lavandier and others and can offer a great possibility to manipulate in detail the different aspects of the stimuli. However, simulations are always dependent on the performance of the respective simulation algorithms and susceptible to artifacts and other problems in signal processing that are not in the focus of evaluation.

Making recordings in real halls with a sound source or sources on stage can be regarded as a good alternative to simulations and being one step closer to a real situation. Of course, the problem of sound rendering is still present, as the recordings have to be reproduced in the laboratory circumstances, but at least, the effects of real acoustics have been this way conserved to some extent depending on the recording and reproduction technique. The use of loudspeakers playing back anechoic recordings ensures that the excitation signal is exactly the same when making recordings in different halls, although, one must consider that one or two sources on stage are hardly representative of a real orchestra. In addition, the recordings are most often performed in empty concert halls when the acoustics are very different from a hall with full occupancy.

Although making "dry" recordings of instruments in anechoic chambers enables playing back the music exactly the same way every time in different halls, it is important to realize that there is one major drawback in these recordings itself. The players and the conductor in a real concert, adjust their playing according to the acoustical feedback and support they get from the hall, and it is this very interaction between the orchestra and the hall what is heard by the listener. In anechoic circumstances there is no such feedback for the player(s), and thus the instruments and the music are possibly played very differently compared with the real situation. Even if it was possible to construct an artificial, full symphony orchestra with all physical parameters (instrument directivities, timbres etc.) taken into account, it would be the very playing with acoustic feedback that would make it different from a real situation.

The terminology used by professionals seems to be quite well established although there still are some discrepancies in the usage of terms, especially in terms related to the aspects of spaciousness in concert halls. However, as the terminology is mainly derived from the physical parameters of sound propagation, it may not be very intuitive for a common concert goer and often results in the problems of semantic

interpretation when used in the perceptual evaluation of acoustics. Acousticians and other experts may be used for the subjective assessments, and arguably they can provide very detailed information about different aspects, but then an issue of generalization to a larger public arises. Thus, there is a need for methods that take into account not only the special and difficult nature of concert hall acoustics but also the issues of psychological and perceptual research. Development of such a methodology requires understanding the sensory analysis techniques which lie at the heart of perceptual psychology and which have been used and elaborated particularly in the food and consumer science. However, in order to incorporate the sensory analysis methods to the field of acoustics research, the development of novel approaches of signal processing and recording must also be considered.

One of the main objectives of this thesis is to present a study where a novel sensory analysis method, the individual vocabulary profiling has been applied in the perceptual evaluation of concert hall acoustics. Before describing this study, the next chapter provides a review of the current state of sensory analysis methods, so that the choice of the applied method can be better understood.

2.4 Summary

This chapter presented an overview of the current state of concert hall acoustics research. First, the main concepts such as reverberation time (RT), early decay time (EDT), definition and clarity (C_{80}), strength factor (G) and loudness, initial time delay gap (ITDG) as well as the measures of spatial impression and spaciousness were highlighted with a historical perspective. Then the various aspects of concert hall architecture were considered.

The emphasis of this chapter was on the previous research on the subjective perception of concert hall acoustics, with descriptions of the main studies and their results. In overall, these studies have applied various methods for the generation of the stimuli ranging from listening to the acoustics in a real concert situation to listening to recorded or processed sound stimuli in laboratory circumtances. Additionally, the methods of perceptual evaluation and quantification of the sensory experience have been diverse. Many of the presented studies have applied interviews and questionnaires (Beranek, Barron, Kahle and Berlin group), while others have used preference tests, paired comparisons and dissimilarity tests (Göttingen group, Lavandier, Soulodre and Bradley). The results in general indicate that the subjective experience and the perception of acoustics is a multidimensional phenomenon, which can be described with attributes such as loudness, reverberance, clarity, intimacy, warmth, envelopment, spaciousness and many others. Alhought, the similarities in the results are clear, there are many discrepancies regarding which and how many attributes are needed to fully describe the perception of acoustics in concert halls. In addition, the relationship between the physical parameters and the perceptual aspects described by the test subjects is still not fully solved.

3 Descriptive analysis and its application to perceptual audio evaluation

The previous chapter presented the main studies of subjective evaluation of concert hall acoustics. These experiments were performed by using a range of different sensory evaluation methods such as preference tests, paired comparisons and dissimilarity tests. This chapter concentrates on sensory evaluation, with the focus being in descriptive analysis and perceptual audio evaluation.

Descriptive analysis is a particular category of sensory evaluation and is often applied in the subjective evaluation of products, especially in the fields of consumer and food science. In general, the goal of descriptive analysis is to characterize the products or the stimuli in terms of perceptual aspects. These tests are often combined with preference judgments in order to evaluate the acceptance of the products and the influencing characteristics.

In the field of audio, descriptive analysis methods have been only quite recently applied in the investigations of the subjective perceptions of sound. At the moment, Lorho's thesis [73] can be regarded as the most detailed discussion about application of descriptive analysis to sound evaluation. There are also a few books such as Bech and Zacharov [11] that cover subjective audio evaluation including the particular topic of descriptive analysis. Outside of the field of audio, sensory evaluation is comprehensively discussed, for example, by Lawless and Heymann [62]. In this chapter, the background and the theory behind sensory evaluation are presented and the main studies where descriptive analysis has been applied to audio evaluation are discussed. Assessor considerations and the statistical methods suitable for data analysis are also covered.

3.1 Background

The foundations of sensory testing are in the experimental psychology, particularly in the branch of psychophysics, which studies relationships between physical stimuli and sensory experience. The first operating characteristic of the sensory system was the notion of just-noticeable-difference (JND), introduced by E. H. Weber in the 19th century. The methods for determining the JND were further elaborated by G. T. Fechner in 1860, who worked out the details of three important sensory test methods: the method of limits, the method of constant stimuli and the method of adjustment or average error. [62]

The method of limits is, for example, the very method used in the traditional audiometry tests, in which the level of sound stimuli is increased or decreased in discrete steps until a change in response is noted. The absolute threshold level, in this case, the hearing threshold level (HRT) is obtained by the average point of change over many trials.

In the method of constant stimuli, the task is to compare the intensity level of the test stimulus against a constant reference level, by responding "greater than" or "less than" to each test item. Also several replications of each intensity level are presented. The results of this test - the percentage of "greater than" responses

- often yield a S-shaped curve, which is commonly called a psychometric function. This function describes the relationship between a parameter of a physical stimulus and the subjective perception. The sensory threshold is usually taken at the point of 50 % on the curve.

In the method of adjustment or average error the test subject is able to control the test stimulus and the task is to match it to the reference. Applications of this method include determination of difference thresholds based on the variability of the subject over many trials, and measuring sensory trade-off relationships, such as, how the duration of a brief tone affects the perception of loudness. [62]

These early test methods of experimental psychology laid the foundations of sensory testing. The methodology first developed by experimental psychologists was adopted to the field of sensory science, and there remain many parallels between the psychophysical and sensory evaluation techniques. Also, it is not surprising that strong interchange between these fields often take place, as they are merely on the opposite sides of the person-stimulus interaction. The sensory psychology is more focused on the person as a research object, while applied sensory evaluation uses people to investigate the properties of the stimulus. It is obvious that these two approaches interleave and cannot be separated, thus, a sensory scientist must be aware of the product development as well as of the factors influencing the subjective perception. [62]

By definition, sensory evaluation is a scientific method used to 1) evoke, 2) measure, 3) analyze and 4) interpret the subjective responses to stimuli, perceived through the senses of sight, smell, taste, touch and hearing. These four activities form the principles and practices of sensory evaluation and each of them must be carefully considered in order to develop a successful experimental design.

First, the stimuli must be prepared and presented so that the possible biasing factors are minimized. For example, concerning sound samples, if we were to evaluate only the effects of spectral variations on the perception of reverberance, it would be necessary to present the stimuli with equal loudness, as loudness variations would probably also influence the perception of reverberance. The presentation order of stimuli may also affect the responses, so, proper randomization may be required.

Next, sensory evaluation applies quantitative methods in which numerical data are collected in order to establish the relationships between the characteristics of the stimuli and subjective perception. These methods are often adopted from the field of behavioral research with the guidelines of application and information about the possible pitfalls and liabilities of these methods. [62]

These two first activities serve the purpose of gathering such data, which can be further analyzed with statistical methods. The data generated by human observers is often highly variable and not all of the sources of variation, such as, mood, motivation, physiological properties, history, familiarity with the stimuli, can be controlled. To evaluate whether the relationships between the subjective responses and the product characteristics present in the data are likely to be real, and not generated merely by change or uncontrolled variation in responses, statistical analysis needs to be applied. There are often many ways for analyzing the same data and different methods may give different and supplementary information about the

phenomenon at hand. By applying multiple analyses the results can be also verified, and, finally, the interpretation of the results may be simplified. [62]

The methods in sensory evaluation can be divided roughly into three categories: discrimination test methods, descriptive analysis methods and acceptance or preference test methods. Before focusing on these different approach, we consider the requirements for the test subjects in terms of the type of information that is desired to obtain with the sensory evaluation. As also stated by Lawless and Heymann [62], the three categories of sensory tests can be further divided into two types: analytical tests and affective tests, which both have very different requirements for the participants.

3.2 Assessor considerations: Analytic vs. affective tests

The first step in performing sensory evaluation is to consider the type of information that is desired to obtain by experiment. Different approaches must be considered in terms of both method and assessor selection, whether the goal is to acquire information about the acceptance of a product or people's preferences, or to evaluate the differences between stimuli or characteristics of stimuli. In fact, according to Lawless and Heymann [62] the central dogma in sensory evaluation is the very distinction between analytic and affective (or hedonic) tests.

Analytical sensory test methods can be divided into discrimination test methods and descriptive analysis methods. In general, the test subjects for these tests are selected on the grounds of having average to good sensory acuity for the critical characteristics of the stimuli to be evaluated [62]. In the audio field, this means that the test subjects should have normal to good hearing and they should be able to listen to and detect variations in sound in terms of the characteristics under evaluation. Depending on the particular experimental design, the subjects are often screened before they are accepted to participate and they may also undergo some training before the listening test is performed. An analytical frame of mind is specially required in the case of descriptive analysis, in which it is essential to be able to put personal preferences and affective reactions aside. In descriptive analysis, the task is to concentrate analytically on the specific characteristics of stimuli, to specify what aspects are present and on what levels of sensory intensity, extent, amount or duration [62].

In many companies, in which sensory evaluation experiments are performed periodically, often a trained and permanent panel of sensory experts is formed. Assessors for these panels are carefully selected based on their reliability, consistency, motivation and availability and they are also subsequently trained for the evaluation tasks. This kind of sensory panel may be regarded as a calibrated measurement instrument, which yields accurate, reliable, consistent and repeatable descriptions of the stimuli. Furthermore, it is common practice to continuously monitor and train the panel to keep up an appropriate performance level. For example, if great interindividual differences are observed between the panelists, there may be a need for further training and discussion about the sensory aspects under evaluation. [62]

In the field of audio, the selection, development, training and monitoring of

subjects and a listening panel have been discussed by several researchers. Bech and Zacharov [11] provide a very extensive overall discussion about this topic. A structured assessor selection procedure has been proposed by Mattila and Zacharov [107] and Isherwood et al. [49] with further discussion by Legarth and Zacharov [65]. The training of assessors has been addressed by Merimaa and Hess [76], Brookes et al. [25] and Neher et al. [81] in the context of the spatial attributes of sound as well as by Quesnel [87] for the evaluation of timbral aspects. Also listening panel considerations have been discussed by, for example, Zacharov and Lorho [106].

In the affective domain, the main objective is often to investigate the acceptability of a product or people's preferences. In contrast to the analytical approach, the preference judgments are made in much more integrative fashion. A stimulus or a product is seen as a whole, and although there might be some specific aspects which draw the attention, the reactions expressed as liking or disliking reflect the overall impression of a stimulus and are often immediate.

On these grounds, naïve test subjects are employed in the acceptability and preference tests as they are effective in rendering the impressions of a stimulus as a whole. Furthermore, preference tests are often conducted with a certain target group in mind, thus, it is important that the participants are part of the population of interest. For example, regarding concert hall acoustics research, the prospective target group would be frequent concert goers or music lovers as they would be likely familiar with the topic and have an understanding about the overall setting in which the stimuli are normally being attended. [62]

There exist a few standards considering the assessors in sensory evaluation. The ISO standards 8586-1 and 8586-2, defined in the context of food industry, are perhaps the clearest and can be adopted to any field in which sensory evaluation methods are applied as discussed by Zacharov and Lorho [106]. These standards define the different assessor types employed in sensory evaluation as well as the development process of sensory assessors from a naive assessor to a specialized expert assessor. The terminology proposed by ISO 8586-1 is reproduced in Table 2.

Considering assessors in terms of type of the test method, a basic principle is to avoid using naïve test subjects in tests, in which an analytical frame of mind is required and expert assessors in preference tests, in which stimuli or products are evaluated in a more integrative fashion. In other words, by employing a small trained group of expert judges, who perform the evaluation in strictly controlled, artificial laboratory circumstances, we may obtain very precise and reliable results, but at the same time, we lose a certain amount of generalizability to the real-world results, what could be obtained by using naive assessors. In every sensory test, there is an amount of trade-off between reliability and precision vs. generalizability and validity to real-life circumstances. [62]

In the current study, this trade-off translates into performing an analytical sensory experiment with a group of naive and inexperienced assessors in laboratory circumstances with short training sessions. That said, one of the main interests in this work is to evaluate the feasibility of such an experiment in the acoustics research. Also, it is interesting to see, if it is, after all, possible to obtainthe accurate and reliable sensory profiles of the halls with naïve assessors.

Table 2: Definition of assessor types in sensory analysis according to ISO standards 8586-1 and 8586-2. Adapted from [62]

Assessor type	**Definition**
Assessor	Any person who is taking part in a sensory test
Naïve assessor	A person who does not meet any particular criterion
Initiated assessor	A person who has already participated in a sensory test
Selected assessor	Assessor chosen for his/her ability to carry out a sensory test
Expert assessor	Selected assessor with a high degree of sensory sensitivity and experience in sensory methodology, who is able to make consistent and repeatable sensory assessments of various products
Specialized expert assessor	Expert assessor who has additional experience as a specialist in the product and/or process and/or marketing, and who is able to perform sensory analysis of the product and evaluate or predict effects of variations relating to raw materials, recipes, processing, storage, aging etc.

Preference tests are not futher discussed in this thesis, as such methods were not applied in this study. Instead, a more detailed discussion about the discrimination tests and the descriptive analysis methods is presented in the next sections.

3.3 Discrimination test methods

Discrimination test methods can be regarded to be the foundation of sensory evaluation as "Discrimination, or the ability to differentiate two stimuli, is after all the fundamental process underlying all other sensory-based responses" [62, p. 141]. These methods have been developed to answer the questions of product similarity before descriptive or affective evaluations are even relevant. It is clear that if two samples cannot be discriminated, there is no point in trying to describe differences between the stimuli.

There are various types of discrimination tests such as paired comparison tests (e.g., same/different tests and 2-alternative forced choice (2-AFC) tests), triangle tests, duo-trio tests, A-Not-A tests, n-alternative forced choice (n-AFC) methods, sorting methods and ABX discrimination tests [62, 19]. A detailed discussion of all of these methods is outside of the scope of this thesis and an interested reader is referred to the books of Lawless and Heymann [62] and Bi [19] for more information

on these methods. Only the triangle test is discussed more thoroughly in the next chapter as it has been applied to evaluate the discrimination abilities of the subjects in the screening phase of this study.

Triangle tests in assessor selection have been previously applied by Lorho [71] in a descriptive analysis experiment and by Legarth and Zacharov [65] in the development of an assessor selection process for multisensory applications. The issues of triangle tests have been addressed by O'Mahony [82]. Otherwise, replications in discrimination tests have been discussed for example by Brockhoff [24], Kunert and Meyners [57], Brockhoff and Schlich [23] and Bayarri et al. [10]. In the audio field, the listening skills and discrimination abilities of candidates have been evaluated by discrimination tests, e.g. by Mattila and Zacharov [107], Isherwood et al. [49], Legarth and Zacharov [65] and Lorho [71]. Data analysis of discrimination tests commonly includes univariate methods (e.g., the analysis of variance (ANOVA)) and statistical testing.

Besides the number or percentage of correct answers, discrimination test results can be analysed to investigate other aspects of the performance of subjects. The most important aspects are the reliability and the consistency of test subjects, of which a good example is provided by Mattila and Zacharov [107]. Moreover, the development and implementation of original test programs have enabled the collection of other performance parameters such as the number of switching between stimuli and response times, as described by Legarth and Zacharov [65] although any formal analysis of these results has not yet been presented. This topic is also elaborated more in the next chapter when the current triangle discrimination test is discussed in detail.

3.4 Descriptive analysis methods

Discrimination tests are usually fairly straightforward and simple, in terms of both test administration and the point of view of the test subject. Although, these tests also require an analytical frame of mind, in a sense, that the preferences and affective reactions are not of interest, the requirement of an analytical mind-set is much more prominent in descriptive analysis (DA) methods, which form the other part of the analytical sensory evaluation domain.

In DA, the task is to identify, describe and quantify the perceptual characteristics of stimuli. In other words, it is the matter of an analytical identification of the perceptual properties of stimuli, which may then be used to compare the stimuli and to differentiate between them. The DA methods are often viewed as the most sophisticated tools in sensory science [62], as they allow the experimenter to obtain a complete description of stimuli and to determine, for example, which perceptual characteristics of a sound are the most influential to the overall listening experience in concert halls.

Elicitation and development of a set of verbal terms to describe perceptions and to evaluate stimuli, is the most popular and widely used method in performing descriptive analysis experiments. These verbal elicitation techniques have been discussed for example by Lawless and Heymann [62] in the field of food industry, and

by Bech and Zacharov [11] in relation to subjective audio evaluation. Also, Lorho [73] has now provided a very comprehensive discussion about the various aspects to be considered in the verbal elicitation process.

As stated by all of these researchers, the underlying assumption in the verbal elicitation techniques is that there is a close connection between a sensation and its verbal counterpart describing the sensation. It is assumed that the (trained) subjects are able to decompose their perception into its constituting elements, create verbal descriptors for these perceptual components and finally, use these terms with quantitative scales to evaluate the intensity or amount of these aspects in the stimuli. Thus, as already discussed previously, these descriptive analysis methods require a highly analytical frame of mind as well as a high level of awareness and sensitivity considering the type of sensation in question.

The origins of verbal descriptive analysis methods are in the field of psychology. Two theories may be highlighted in this respect, namely the semantic differential developed by Osgood (1952) [83] and the Repertory Grid developed by Kelly (1955) [55]. Osgood presented the semantic differential as a general method of measuring the connotative meanings of concepts. Basically, his method employs seven-step bipolar adjectival scales which are presented to the test subject with a task to indicate the direction and the intensity of the association regarding each specific concept item. Kelly's theory about personal constructs has many similarities with the semantic differential approach and is discussed in detail later in this chapter.

There are two main approaches in performing verbal descriptive analysis: the traditional consensus vocabulary profiling, referred as CVP, and more recently developed individual vocabulary profiling, referred as IVP. Although, there are many methodological similarities between these two approaches, the differences are considerable in terms of practicality, resource and time requirements and data analysis. The main difference can be however described in terms of elicitation and development of the verbal descriptors. In CVP, the verbal descriptive terms are developed with a panel of assessors and the panel is trained by group discussions to use the terms in a similar and consistent way while in IVP the verbal descriptors are elicited and developed individually by each assessor and no group meetings are conducted.

Lorho [73] may perhaps now be regarded as one of the main contributors to the descriptive analysis of audio with his thesis on the perceived quality evaluation of sound reproduced over headphones. In his thesis, Lorho has applied both CVP and IVP in the evaluation of sounds listened over headphones. He also provides a detailed comparison of these two methods. Other studies employing a descriptive analysis approach in subjective sound evaluation have been conducted by Berg and Rumsey [18], Zacharov and Koivuniemi [105], Guastavino and Katz [45], Choisel and Wickelmaier [28], Kim and Martens [56] as well as Lorho [72, 71].

There are also other descriptive analysis methods, namely non-verbal and indirect elicitation methods, which have been discussed shortly by Lorho [73] as well as by Bech and Zacharov [11]. Instead of eliciting verbal descriptive attributes for the evaluation of stimuli, the non-verbal and indirect elicitation techniques make use of other forms of expression, such as drawing (e.g., Ford et al. [79] and Lokki et al. [67]) or hand gestures (e.g., Lokki et al. [70]), the indications of similarity or

dissimilarity (see multidimensional scaling method, MDS, e.g., Bonebright [20]) and sorting methods (e.g., Cartier et al. [27]). The overall descriptions are then obtained by multivariate data analysis.

The focus in this thesis is on the direct verbal elicitation methods, particularly on individual vocabulary profiling. A discussion about the differences between CVP and IVP is presented in the next sections as these methods are tightly related. Introduction to data analysis techniques, particularly multivariate methods employed in IVP, are also presented in this chapter. Non-verbal and indirect elicitation methods are out of the scope of this thesis and are not further discussed. In the following, the general outlines of consensus vocabulary methods are presented before individual vocabulary profiling methods are discussed more in detail.

3.4.1 Consensus vocabulary methods

On historical grounds, describing the sensory characteristics of stimuli and sensory evaluation was a task for one or a limited group of experts, who had had extensive training or experience of the different aspects of the stimuli. However, the various disadvantages (e.g., generalizability to a larger context, objectivity and practical limitations) in this approach motivated the development of more sophisticated sensory evaluation techniques, which apply the same basic ideas of the earlier methods, but with a formal structure and scientific robustness. Consensus vocabulary profiling (CVP) methods employ a group of assessors to develop a common terminology to describe and evaluate stimuli. The assessors are trained for the task and the results are obtained with statistical data analysis techniques to alleviate the issues of interpretation.

There are several approaches to CVP, differing in terms of the elicitation and development of attributes, the training of assessors, scale usage, administration etc. Four main CVP techniques, which have been developed primally in the context of food industry, are (in chronological order): The Flavour Profile Method ™[26] (1950), The Texture Profile Method ™[96] (1963), The Quantitative Descriptive Analysis ™[95] (1974) and The Spectrum ™(1970s). Detailed descriptions of these methods are not provided here but are presented, for example, by Lawless and Heymann [62], and Lorho [73]. However, the general outlines of these methods are basically the same and can be described with three steps according to Lorho [73]: (1) panel selection, (2) the consensus vocabulary generation and (3) panel training.

The assessor selection has been already discussed previously in 3.2. After a panel (usually 6 to 20 assessors) is formed, the next step is to develop a common vocabulary to be used in the evaluation of the sensory characteristics of the stimuli. Term "concept alignment" is often used to refer to this process of finding an agreement of the attributes between the panelists. This often includes several phases, where assessors first familiarize with the stimuli, identify and describe the sensory properties individually and in groups, and finally establish an agreement and a single list of attributes for the whole panel. In addition, they are often required to provide definitions for the attributes, choose the intensity scales to be used in the evaluation and select or create physical references for the attributes.

The last step before the formal evaluation is to train the panel to ensure that the assessors have achieved an agreement of the terms, and are able to use them consistently and reliably. As stated by Labbe et al. [59], training is arguably a critical step in obtaining reliable sensory profiles and it has a significant influence on the quality of sensory evaluations. It is notable that, vocabulary development and training phases may require more than 10 separate group sessions.

The data from CVP experiments can be analysed with relatively simple and robust statistical methods. ANOVA and other general linear models are always helpful in the analysis although the overall sensory profiles of stimuli are usually obtained with multivariate methods such as, Generalized Procrustes Analysis, GPA, and (Hierarchical) Multiple Factor Analysis, (H)MFA. These methods can be used to reveal the latent sensory space, by which the sensory properties of the stimuli can be interpreted.

Besides these product profiles, experimenters are often interested in the performance of individual assessors as well as of the whole panel. The different aspects of assessor and panel performance in the consensus vocabulary profiling experiments are well discussed by Lorho [73], the three important aspects being repeatability, agreement and discrimination. Univariate and multivariate methods may be both used to investigate the panelist and the panel performance and they provide different perspectives on these aspects. Data analysis techniques will be attended in detail later in this chapter.

Consensus vocabulary techniques are still perhaps the most applied methods in the field of sensory science as they result in a detailed and accurate descriptions of the stimuli. However, the panel forming, vocabulary development and training requires much time and commitment as well as solving many practical issues. Considering the tight schedules often present e.g., in product development, these resource requirements may be too heavy for the company or there is just no time for detailed product evaluations. Thus, reducing the requirements of CVP can be seen as one of the key questions in the modern sensory evaluation practises, and as a motivation for many developments in the methodology of sensory science. The emergence of IVP methods can be maybe regarded as the main advancement in this respect, eliminating the need for concept alignment by group discussions. The IVP methods and data analysis considerations are discussed in the next sections.

3.4.2 Individual vocabulary profiling methods

Individual vocabulary profiling methods in general consist of each assessor developing an individual set of descriptive attributes which they apply to evaluate and compare stimuli or products. These methods do not require any group meetings between the assessors and any form of concept aligment is unnecessary. Thus, this approach is considered much faster and require less resources than the conventional consensus vocabulary profiling, but the same principles and requirements still apply and need to be considered.

It was stated that in verbal elicitation techniques, which include both CVP and IVP methods, the underlying assumption is that assessors are able to identify

separate perceptual properties constituting their overall experience of stimuli. This assumption may not be apparent in the CVP as extensive group discussions and training ensure that assessors understand the meanings and perceptual properties which the terms refer to. However, in IVP, the deficiency of this assumption may be much more prominent when the experimenters must rely on an individual assessor's ability to produce relevant, descriptive and hopefully also discriminative attributes. Although, the experimenter may ensure that the attributes are non-affective and relevant by discussing the matter with the assessor, ultimately only the assessor knows in detail which perceptual aspects the attributes relate to.

Although assessors are usually screened and selected for experiments of this kind, often they are naive and inexperienced regarding sensory evaluation. However, the possibility to use inexperienced assessors who can be quickly trained, may be regarded as one of the advantages of IVP. It is commonly accepted that test subjects in general are more reliable and consistent when they are able to use their own attributes for the evaluation. However, the elicitation and development of attributes is not a simple process. It requires good perception skills from the test subject and careful planning from the experimenter.

Four main techniques of individual vocabulary evaluation are The Repertory Grid Technique (RGT) [55], Free-Choice Profiling (FCP) [103], The Flash Profile (FP) [94] and Individual Vocabulary Profiling (IVP) [73, 72, 71]. In the audio domain, these methods have been discussed in detail by Bech and Zacharov [11], and Lorho [73], who also provides a very comprehensive and detailed discussion about the statistical analysis of IVP data. These four methods are described in detail in the following.

The Repertory Grid Technique

The Repertory Grid Technique (RGT) was developed in the field of psychology by George Kelly [55] in the 1950's. The RGT is based on Kelly's theory about personal constructs, on an idea, that a person's understanding, perception and interpretation of the surrounding world are made up by a system of constructs. These constructs are thought to be dichotomous with extreme points, e.g., "loud -soft" or "clear - muddy". The RGT was originally developed to investigate and reveal these internal constructs, particularly constructs related to the aspects of personality. However, now it has been also applied in the sensory evaluation domain to elicit characteristics of stimuli. For example, in the audio field the RGT has been employed by Berg and Rumsey [17], in order to identify the spatial attributes of sounds.

The RGT consists of presenting the subject sets of three samples (i.e. triads) with a task to indicate not only which sample differs the most from the other two, but also to describe the characteristics in which way two samples are similar and different from third one. After all the possible triads have been presented, the elicited descriptions are used to produce a "grid" of word pairs, each pair representing one construct with the descriptors as extreme points. This grid is then applied in the evaluation of all the samples. Samples are rated with each construct on a scale that has been determined by the experimenter. According to Bech and Zacharov [11],

the possibilities for the rating procedure include, dichotomisation (i.e., which one of two words best describes the sample), use of a category or a continuous scale or ranking. Depending on the rating technique, results are usually analysed with statistical methods such as principal component analysis, factor analysis and cluster analysis. Berg [16] has also developed software called OPAQUE in order to facilitate the generation of the grid, rating procedure as well as analysis and presentation of the results.

Free-Choice Profiling

Free-Choice Profiling (FCP) method was introduced by British sensory scientists in the 1980's as an alternative approach to conventional consensus vocabulary techniques. On early on, it was applied in the evaluation of various food products, e.g., wines by Williams and Langron [103], coffees by Williams and Arnold [102], cheese by Marshall and Kirby [74] and chocolate by McEwan et al. [75].

In FCP, assessors develop their own individual lists of attributes, and very little training is performed before the evaluation to reduce the time requirement of the experiment. Thus, often it is desired to use assessors who already have some experience in sensory evaluation although FCP have also been applied to consumer research (see Jack and Piggott [51]) with naive test subjects. The stimuli are usually presented separately and evaluated with all attributes at a time and the resulting data is commonly analysed with multivariate analysis methods in order to obtain a consensus profile from the individual sensory scores.

The Flash Profile

A more recently developed IV technique is the Flash Profile (FP) introduced by Sieffermann [94] in the field of food industry. In the audio field FP has been reviewed by Bech and Zacharov [11] and Lorho [73]. There are also several publications, e.g. Dairou and Sieffermann [29] and Delarue and Sieffermann [31] where FP has been compared with more conventional sensory evaluation techniques. The results are promising in terms of similarity of the sensory profiles between FP and more traditional methods as well as the time requirement of the FP, but there are still certain uncertainties that have been noted by these authors. For example, the interpretations of the sensory profiles have been reported being more difficult in the FP than in a conventional CVP approach.

In the FP, the individual elicitation approach of FCP is combined with a comparative evaluation of all samples. The comparative evaluation technique has been argued to remove the requirement of familiarization and individual training phases and therefore, to further reduce the time requirement of the experiment. However, the lack of familiarization and training means that usually only assessors who are already familiar with the sensory analysis are employed to ensure that the attributes are not affective. Also, the comparative evaluation puts some limitations on this method, as a simultaneous presentation of the samples is required and only a relatively small number of samples may be evaluated at a time. Still, the main advantage is that the results may be obtained in as few as one to three sessions with FP. On

a final note, instead of being an alternative, FP is often regarded as a supplement to conventional methods and a quick way to conduct a preliminary study of stimuli before a more thorough investigation is performed (see Tarea et al. [98]).

The Individual Vocabulary Profiling

The most recent advancement in IV techniques has been in the field of sound evaluation with the Individual Vocabulary Profiling method developed by Lorho [73, 72, 71]. Lorho's approach combines features from the RGT, the FCP and the FP and he has described this method as "a relatively efficient sensory profiling procedure tailored for sensory testing with inexperienced assessors" ([73] p. 135).

IVP is based on the comparative evaluation approach presented in the FP, what arguably facilitates the attribute generation process and improves the discrimination of sensory properties. A similar feature with the RGT is the diadic and triadic presentation of stimuli which is applied in the elicitation phase of descriptors. Lorho points out that this way the elicitation is more structured and it helps naive and inexperienced assessors to achieve a set of attributes which is descriptive and discriminative. Also, additional effort is put to gather attribute definitions, which provide valuable semantic information and can be used in the interpretation of the results. In contrast to the FP, a training phase is also included to further compensate for the inexperience of assessors. The rating of stimuli is performed with continuous intensity scales which quantify the amount of the perceptual characteristics present in the stimulus set. In accordance with the other IV methods, the analysis of the results can be performed with a range of statistical analysis techniques, which are discussed in the next section. Lorho [73] has also proposed a semi-automated system for the design and administration of an IVP experiment including a GuineaPig 3 user interface as well as a MATLAB routine to handle and analyse the attribute rating data. An overview of the IVP is illustrated in the Figure 2 according to Lorho [73].

The sensory evaluation method employed in the current study is based on Lorho's IVP method presented above. However, some modifications have been made in order to match with the resources at hand and other practical considerations. These are described in detail in the next chapter with a full description of the implementation of IVP experiment to concert hall acoustics evaluation.

3.5 The analysis of individual vocabulary profiling data

An individual vocabulary profiling experiment can be conducted with various methods as discussed in the previous section. Although, there are some differences in the experimental design between the different methods, the analysis of the data is essentially similar.

The difficulty of data analysis arise from the complexity and the amount of the data. Each assessor evaluates the samples with his/her own set of descriptive attributes what results in a set of individual data matrices. Although, it is already interesting to investigate these individual matrices in isolation, and to obtain the

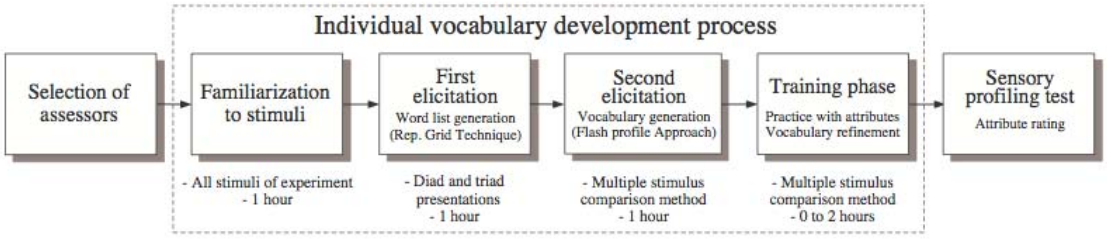


Figure 2: Steps of the individual vocabulary development procedure after Lorho [73]. A total of 3 to 5 hours is needed for this process depending if a training phase is included or not.

individual sensory profiles, more commonly the experimenter's interest lies in obtaining the overall descriptions of the products or samples incorporating all data of all the assessors. However, in order to get these averaged sensory profiles, often referred also as product spaces, one must consider multivariate data analysis techniques, which involve heavy mathematical manipulations of the data at hand. This chapter will introduce the most commonly used analysis methods used in the IVP experiments, such as Principal Component Analysis (PCA), Multiple Factor Analysis (MFA) and Generalized Procrustes Analysis (GPA).

The analysis of the data can be divided into two parts: 1) obtaining the sensory profiles and descriptions of the products and 2) analysing the reliability of the results, that is, the performance of individual assessors as well as the whole panel. Considering the current study, the focus in this thesis is on the individual sensory profiles analysed with PCA and *beta*-coefficient, which is a measure of dimensionality of a data matrix. This data analysis is presented in the next chapter. In this section, an overview of the relevant statistical analysis methods is presented.

3.5.1 Properties of IVP data

Considering the qualitative aspects of the IVP data, there is a large amount of information embedded in the attributes and their definitions. The attributes and their definitions are interesting as such, but a more analytical approach can be useful in comparing the individuals and their attributes. As Lorho [73] discusses, one feasible approach is proposed by Gaines and Shaw [35] although their domain of application is more general than in the case of IVP. These researchers used RGT with two expert assessors to elicit 'distinctions', which can be characterized as bipolar constructs and 'terms' to define the end points of the bipolar constructs. According to Lorho, the 'distinction' can be regarded as equivalent of an 'attribute' developed by an assessor in the case of IVP. When these individual conceptual systems are compared, four different scenarios arise: consensus, conflict, correspondence and contrast as illustrated in the Figure 3.

Consensus is the case when two individuals define the same distinction with the same terms. Conflict in turn arises when the same terms are used for different dis-

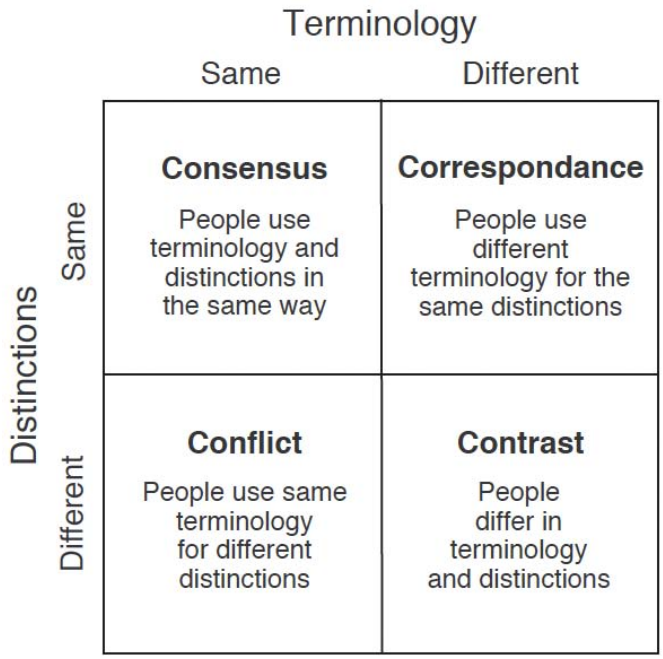


Figure 3: Terminology of different scenarios after to Gaines and Shaw [35]

tinctions. Correspondence happens when the same distinction is defined with different terms and contrast is the case when individuals do not use the same vocabulary at all, that is, they have different distinctions with different terms defining them. Lorho notes, that the data structure of IVP is very suitable for being investigated with this approach as the IVP data consist of individual attributes with definitions and their ratings. Besides, when analysing the quantitative rating data and comparing the individual profiles, e.g. with PCA, it is possible to use this framework for better identify and understand the interactions between the individuals.

The quantitative rating data of an IVP experiment are a multidimensional data sets comprising of N matrices (one for each assessor), X rows (samples) and Y_N columns (attributes). If there is multiple sets of stimuli, as is the case in this study and, for example, in [72, 71], there is a set of matrices for each assessor. Note, that in IVP also the number of attributes, i.e., Y can differ between assessors. This data structure is represented in the Figure 4. It is clear that only multivariate data analysis techniques can be considered due to the special structure of the data. An overview of the most common methods is presented next.

3.5.2 From individual profiles to a combined product space

Considering the structure of the IVP data, a natural approach to analysis is to first investigate the properties of each individual data set separately before an overall analysis is performed. One helpful tool in analysing the relationships between the attributes of an individual sensory profile as well as investigating the complexity of

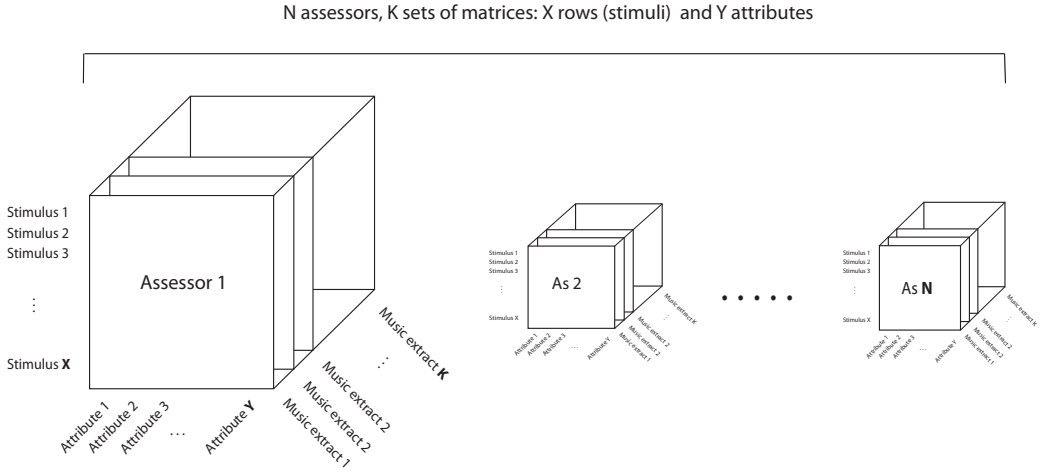


Figure 4: Representation of the structure of IVP data

a sensory profile is the Principal Component Analysis (PCA). The complexity of a sensory profile can be also investigated with a *beta*-coefficient proposed in the article by Shlich [90]. These methods are discussed in following.

Individual sensory profiles with Principal Component Analysis (PCA)

Principal Component Analysis is thoroughly discussed by Jolliffe [52], and he summarizes the main idea of PCA as follows: "The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. " ([52], p. 1))

By PCA, it is possible to analyse the structure of the individual configuration and investigate the relationships between attributes and samples in the multidimensional space spanned by the principal components. The complexity of the sensory profile can be evaluated with the variation explained by each principal component. This information is also reflected by the correlations of the attributes in these dimensions. In this context, also a very useful term defined by Lorho [73] is *perceptual direction*, which can be used to describe the sensory pattern identified with an attribute or a group of attributes in this latent space defined by several dimensions.

The objective of the attribute elicitation and development process is to define a set of attributes, which is both descriptive and contains as much information as possible regarding the samples. In other words, the attributes should relate to different perceptual properties, that is, to properties according to which the samples can be well discriminated. If the attribute development process has been

successful, it is reasonable to expect attributes which discriminate the samples in multiple perceptual directions. Thus, PCA can be used to verify the structure of an individual configuration and to pinpoint attributes that are interrelated and do not offer information on separate perceptual directions.

The high correlation between attributes may be due to many reasons. First, it is possible that the attributes are in fact related to different perceptual aspects but the variation in these aspects is in the same direction, that is, the aspects are heavily related with one another. It is also possible that the different attributes are in fact describing the same perceptual aspect, or there is yet another charasteristic that has dominated the perception of these aspects resulting in a high correlation. In all of these cases, it can be argued that the assessor has had some difficulty in developing attributes by which different information about the samples can be extracted. Whether these difficulties are due to the properties of the samples or to the abilities of the assessor can hardly be answered although indications to one way or another can be obtained by performing comparisons between the individual configurations. In the next chapter, these ideas are discussed more with the examples of the current study.

One way to investigate the complexity of a sensory profile is to look into the explained variation of principal components (PCs) and see how many components are explaining most of the total variance. Most commonly the eigenvalues are used for this, so that the results are interpreted only for the PCs that have an eigenvalue greater than one. However, in addition to PCA another method in order to evaluate the complexity of an individual configuration is the β-coefficient proposed by Schilch [90].

β-coefficient as a measure of complexity of a sensory profile

The classical problem of PCA is to decide how many principal components are retained in the analysis. The different approaches are discussed for example by Jolliffe [52]. However, instead of investigating the principal components for the dimensionality estimation, Schilch [90] proposes a β-coefficient as an estimator of the dimensionality of an individual sample space. Because this measure has been less used in the previous studies, the main features of the mathematical derivation are presented here according to Schilch. In the following, it is assumed that the data set is centered for the assessor.

Let X_i be a data set of an assessor, with n rows (the number of samples) and p columns (the number of attributes). The association matrix W is defined as:

$$W = X_i X_i' \tag{2}$$

Note that this matrix contains the full information about the multidimensional relationships between the different samples regarding the assessor i. β -coefficient is calculated from the association matrix as follows:

$$\beta = \frac{(trace(W_i))^2}{trace(W_i^2)} \tag{3}$$

where *trace* refers to the sum of the diagonal elements of a matrix. Without going into further details about the aspects of mathematical derivation, there are two important properties of the β -coefficient what give clear indications why it can be used as a dimensionality estimator. Firstly, the lowest dimensionality (a single axis) is obtained when the attributes are fully correlated and secondly, the highest dimesionality is obtained when there is no correlations between the attributes at all. In other words, β -coefficient can vary from 1 to $P_i = \min(\text{n - 1}, p_i)$ where p_i is the number of attributes of the assessor. Note, that it is probable that there is at least some correlation between different attributes, so the highest dimensionality is not likely achieved. Schilch also highlights that alhough the β-coefficient can give an indication of the number of ideal attributes that would be sufficient to describe the sample differences, it should not be regarded as being an exact truth about the dimensions implicated in the sensory evaluation. β -coefficient can be understood as a measure of complexity of an individual sensory profile and used as for the analytical comparison of dimensionality of profiles.

The analysis of the individual sensory profiles serves well as a first step in the analysis process, but usually the objective is to obtain an overall sensory profile, which may be generalized to represent the perceptions and opinions of a larger population. PCA is useful to investigate an individual data set of one set of attributes, but as different assessors have their own sets of attributes, the global analysis needs to be performed with other multivariate methods, which can account for this discrepancy in the data set. An average sensory profile can be obtained with methods such as Multiple Factor Analysis (MFA) and Generalized Prorustes Analysis (GPA) and although these are not employed in this thesis, they are presented next, because they are an essential part of descriptive analysis.

Additionally, a very useful tool for investigating the relationships of the attributes of all assessors is Hierarchical Cluster Analysis (HCA), which is applied to make a grouping of the attributes according to their rating data. This information can often be useful in order to verify the semantic interpretation of the attributes and to simplify the interpretation of the overall sensory profile. A detailed discussion of HCA is however out of the scope of this thesis

Multiple Factor Analysis

Multiple Factor Analysis (MFA) is discussed in detail, e.g., by Escofier and Pagès [34]. As already stated above, MFA is one proper method in order to obtain an overall sensory profile in the framework of IVP. MFA analyzes several sets of data where the same individual are described with different groups of variables. The variables between the groups may be of numerical or categorical type, but they should be of the same type inside one group. Additionally, the number of variables in each group can be different. The main features of MFA are discussed shortly in the following according to Escofier and Pagès [34], Pagès [84] and Lorho [73].

In order to obtain an integrated picture of the results, the first step in MFA is to

make the groups of variables comparable. This is performed because otherwise the group of variables with the strongest structure would dominate the average space. Balancing the groups is achieved by weighting each set X_i of centered variables by the inverse of the first eigenvalue of the variace-covariance matrix $X_i'X_i$, which is also called a first singular value. In practical terms, this step can be performed by applying a PCA in each group of variables, as the first singular value is the square root of the first eigenvalue of the PCA. After, balancing the data sets they can be concatenated and submitted to another PCA, which results in an overall sensory profile.

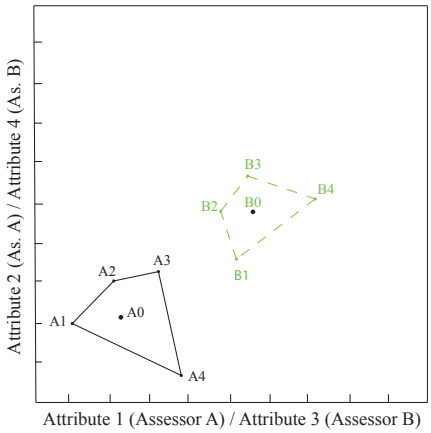
A special situation arises when there are multiple datasets for each individual. This occurs, for example, in an experimental design in which the assessors evaluate various signal processing algorithms with their own attributes, but with various extracts of music or sound. Another special case occurs, when the experimenter wishes to analyse and compare the outcome obtained with several trained sensory panels with the results of an untrained panel or to compare the sensory results with the results from physical measurements. In these situations, a Hierarchical Multiple Factor Analysis(HMFA) is the proper method as discussed by Dien and Pagès [64]. HMFA account for the hierarchical structure in the data and provides outputs that can be interpreted on an overall level as well on the different levels of the hierarchy. Various visualization techniques are usually helpful in order to interpret the results.

Generalized Procrustes Analysis

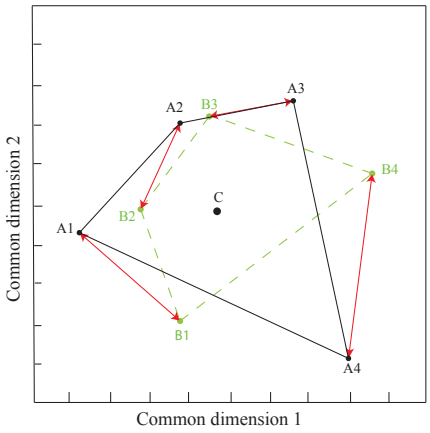
The generalised procrustes analysis is one of the most commonly used methods of data analysis in sensory experiments. It is applicable to both conventional consensus panel data and to data from individual vocabulary profiling. In this section GPA is discussed according to Dijksterhus [32], and the main features of the analysis are outlined without going into the formal definitions. A detailed mathematical derivation of GPA is given by Gower [36] and a summary of this method and its application to IVP data is discussed, for example, by Lorho [73].

Originally, the procrustes analysis was developed by Hurley and Cattell [47] in 1962 as a process of matching two matrices of N objects by M variables. After in 1975, this was generalised by Gower [36] enabling the analysis of multiple data sets with the possibility of a differing number of variables (columns) between the sets. As Lorho [73] describes, the procrustes analysis approach is maybe best understood by considering a geometrical configuration of N points (objects) that lie in M (variables) dimensional space for a number of K assessors. The objective of GPA is to minimize the distances between the points of different configurations by performing a set of geometrical transformations on the configurations. These transformations include translation or shifting, rotation/reflection and isotropic scaling (stretching or shrinking) as illustrated in Figure 5. An important constraint in performing these transformations is that the relative distance between the objects in one configuration has to be preserved.

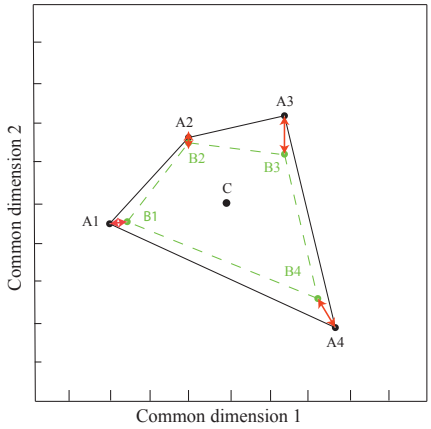
Translation/shifting is performed in order to correct the so-called level effect, which is also known as the assessor main effect in the analysis of variance. Level



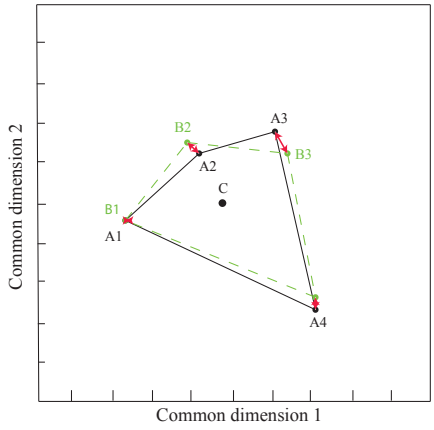
(a) Original configurations of two assessors A and B.



(b) Centred configurations after translation and shifting. Red lines represent the distances between the product points.



(c) Configurations after centering and rotating.



(d) Configurations after centering, rotating and isotropic scaling.

Figure 5: Illustration of the data transformations applied in GPA. Adapted from [90]. Note that in this illustration the assessors may have used different attributes (as seen on the axes of Figure 5a) but they are thought to be perceptually related.

effect manifest in different average scoring positions of assessors. For example, on a line scale from 0 to 100, one assessor may use a range from 20 to 60 while another uses a range from 60 to 100. However, it is possible that these assessors would agree on the perceptual aspects with one another if they had used the scaled similarly. In geometric terms, this level effect is removed by translating the entire configurations of assessors so that the centre points of the configurations coincide with each other.

In mathematical terms, this is known as column centering.

Rotation/reflection of configurations accounts for the possibility that assessors have different interpretations of attributes or that the attributes are different all together. This is performed by rotating the entire configurations so that the N object-points of the different data sets are in agreement with each other. Additionally, also reflecting the configurations in a particular dimension is possible if necessary. Mathematically, this operation is represented in a rotation matrix H_k for the assessor k.

Isotropic scaling, i.e., strecthing or shrinking is performed to account for a so-called range effect. Range effect occurs for example, when one assessor evaluates the objects with a scale range from 30 to 70 and another uses the range from 10 to 90. Note that although the range effect is similar to the level effect discussed above, these are different effects which are both caused by the differing scaling behaviour of assessors. Sometimes when it is possible, the range effect as well as the level effect can be somewhat prevented by advising the assessors to evaluate the objects with the full scale. In GPA, the scaling factors are presented by a number p_k, which is larger than 1 when the configuration is stretched and between 0 and 1 when it is shrunk.

After the distances between the corresponding points of the configurations are minimized with the aforementioned transformations, they can be interpreted by the means of ANOVA and PCA. There is various aspects that can be extracted with these methods. Note that, a group average i.e. the average of the corresponding points between the configurations is used in order the represent the results and analyse the relations between individual assessors and the average configuration.

The 'variances' can be obtained by squaring the distances between the corresponding points of different configurations. By adding them, we obtain an overall measure of loss and by comparing it with the squared distances before the GPA, we get a measure of the loss that cannot be modelled with GPA. On the other hand, the fit of the model is the complement of the percentage loss to 100 percent. Additionally, by adding the variances over N objects per assessor, a measure of agreement between an individual assessor and the group average is obtained. Moreover, if these variances are added over K assessors for each product or sample, a measure of the amount of agreement among the assessors for a particular product can be obtained. This way outlying assessors and products can be detected.

According to Gower [36], the transformations are performed in the highest dimensionality, that is, 100 percent of the data are used throughout the analysis. This results also in a high dimensional optimal solution, which is difficult to illustrate and interpret properly. PCA is a convenient tool for reducing the dimensionality and represent the results in a low dimensional space. PCA is applied to the group average configuration and the results can be plotted, for example, into two-dimensional space. The percentages of the explained variances of the PCA dimensions are useful in investigating the dimensionality of the solution and can indicate the number of dimensions that are needed for the interpretation of the results.

The original variables - attributes - can be also illustrated in the group average space by two means. It is possible to use the coordinates of the rotation matrices,

which are called the loadings of variables, or the correlations between the original variables and dimensions of the group average space. They both infer the same information, so it can be regarded as matter of taste, which one is used in the analysis.

In addition to sensory profiles, one interesting outcome of the GPA is a measure called "RV coefficient". RV coefficient is a generalized Pearson correlation coefficient between two matrices and can be used as a measure of the level of similarity of two sensory profiles. Thus, as Lorho [73] points out, it can be used to evaluate both the repeatability (i.e. the similarity of repeated sensory assessments) and agreement (i.e. the similarity of two different sensory profiles) which are used to evaluate assessor and panel performance (see the next section 3.6). The RV coefficient is thoroughly discussed by Shlich [90].

There has been discussion about the statistical significance of the GPA results but there still not exist a formal test for significance. Perhaps the most common approach to address this matter is the permutation tests (see e.g. Wakeling et al. [101] and Xiong [104]) which, make use of the approximations of a permutation distribution due to the fact that permutation tests can be very time consuming. GPA is implemented in FactomineR [63] package for R [48] statistical language and environment.

3.6 Assessor and panel performance

In the previous sections, the most common methods are presented for analysing the results and investigating the sensory profiles on an individual and panel levels. In this section, a more detailed look into the considerations of the reliability and accuracy of the results is given. In a framework of sensory analysis, an individual panelist or a panel as a whole can be regarded as a measurement instrument, from which, reliable and accurate results are expected. The reliability and accuracy in this context are commonly addressed throught concepts of repeatability, agreement and discrimination. These aspects are extensively discussed, for example, by Pineau [86] and Lorho [73].

- Repeatability

  Repeatability can be only considered in the case where replicated assessments are included in the experiment. It is the matter of the closeness of the results of the same assessor or the same panel in a set of repeated measurements. Regarding an IVP experiment, the repeatibility is usually measured on the level of single attribute by the means of analysis of variance (ANOVA), but it can be also measured on the level of assessor or even the whole panel.

- Agreement

  Agreement relates to the inter-agreement between different measurement systems, which are in this context single assessors or whole panels. In this thesis, agreement is discussed in terms of inter-agreement between panelists. Moreover, keeping in mind the special structure of the IVP data, i.e. that the

attributes are individual, agreement can not be investigated in the univariate domain. It can be evaluated only with multivariate analysis methods and including the complete sets of the assessors' data.

- Discrimination

  Discrimination relates to the ability to perceive differences in the sample set as discussed previously with the PCA. As Lorho [73] describes, discrimination has a particular disposition in the sensory analysis as an "accepted true value" is not available. While two former criteria are independent, discrimination, in turn, is not. Considering a single panelist, discrimination is related to the repeatability as it is difficult to achieve a good level discrimination if the repeatability is poor. Then, at a panel level, disrcimination depends also on the level of agreement between the panelists. Thus, if the agreement between assessors is poor, the discrimination remains poor in the panel level, even if all the assessors show a high level of repeatability and discrimination individually.

The issues of panel and panelist performance have been discussed by several authors with strategies for the evaluation of these matters. Perhaps the most thorough discussion is provided by Pineau [86], who compares different approaches by applying data from a large number of different sensory experiments. She also discusses these matters with a longitudinal perspective. Lorho [73], also addresses the performance issues in both CV and IV experiments and points out the differences in performance evaluation in these cases. In this thesis, the focus is on the panelist and panel performance evaluation regarding especially IVP experiments, although, the general ideology is very much the same also for a consensus panel. The main discrepancy is that in IVP, agreement can not be evaluated in a univariate domain, because the attributes are not commensurable and the methods utilising mean scores of the panel are not applicable.

Considering the panelist performance on the level of a single attribute, the evaluation of the repeatability and discrimination is the same for the CV and IV approaches. As Naes and Solheim [80] describe, it is possible to use the one-way ANOVA model to evaluate both repeatability and discrimination in the IVP experiments. The repeatability of a single attribute can be addressed with the mean square error (MSE) of the repeated scores, which represent the residual variance of the model. Discrimination in turn is related to the "product" or "treatment" effect whose significance is measured by the F-ratio or its associated p-value. In addition Tomic et al. [99] present several visualization techniques for an easier presentation and interpretation of the performance evaluation.

Although, the evaluation of performance of separate attributes gives a very detailed information about the assessors' performance, it may also result in a overwhelming number of analyse considering that an IVP experiment often yields a large number (more than 100) of attributes. Thus, it is often more feasible to analyse the panelist as well as the panel performance in the multivariate domain, when also the level of agreement between the panellists can be addressed this way. In addition, the analysis of the performance can be carried out parallel with the analysis of the sensory characteristics of the products.

The differences between the univariate and multivariate perspectives in the performance evaluation are well discussed by Schlich et al. [91] as well as Lorho [73]. They illustrate the interpretation of the performance in these domains with a following example. Let's consider a simplified sensory data consisting of two assessors, two attributes, three products and the scores from three repeated evaluations per product. In Figure 6 the corners of the triangles represent the replicated assessments and the letters A, B, C represent the products in question. Two assessors are presented with different colors i.e. solid blue line for assessor "blue" and dashed red line for assessor "red".

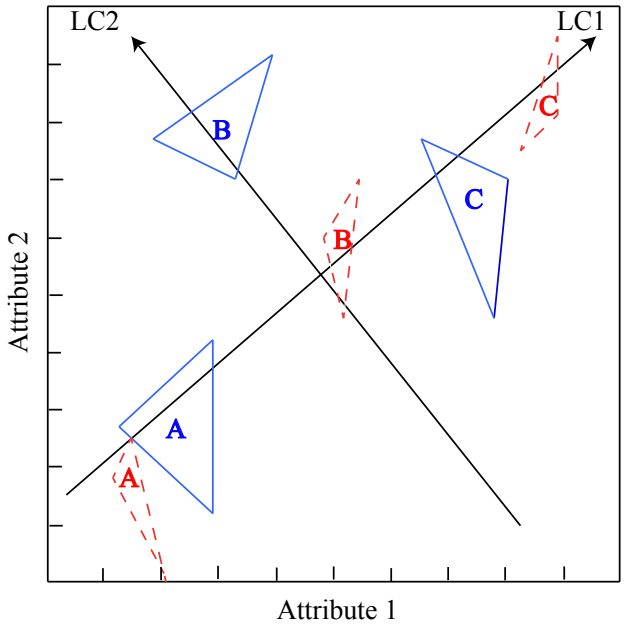


Figure 6: An illustration of panelist performance evaluation with a simulated sensory data consisting of two assessors, three products and three replicates. Illustration is adapted from [91] and [73].

In the univariate domain each of the attributes is investigated separately and the results are projected into the original axes. Regarding Figure 6, the following information can be extracted with this viewpoint. First, it seems that the assessor "blue" is less repeatable on both attibutes than assessor "red". Importantly, because of this lower repeatability, assessor "blue" is also less discriminative on both attributes. Considering the assessor "red", he is less repeatable for the attribute two than the attribute one and, thus, also less discrimative in this respect. Considering agreement between the assessors, they have agreed more on the attribute one than attribute two although in the case of IVP, this could not be inferred in the univariate domain.

In the multivariate perspective, the results are considered directly in two-dimensional space. The principal directions of variation are presented by two arrows

in Figure 6 and referred as the latent components LC1 and LC2 respectively. Usually, the multidimensional interpretation would be done with coordinate axes presenting the latent components instead of the variables themselves, but this example can be well used to also illustrate the multidimensional approach. First, it is seen that the sensory configurations of these assessors are different. Apparently, the configuration of assessor "blue" is bidimensional while it appears to be one dimensional for assessor "red". The higher product distances on the first latent component indicate that the assessor "red" has been more discriminative in this respect, but in spite of the poor repeatability of assessor "blue", he has been somewhat more discriminative in the second latent component direction. These two aspects are differentiated by Shlich et al. [91] as the strength of product discrimination and the dimensionality of product discrimination. While the dimensionality of discrimination is greater for assessor "blue" than the assessor "red", it seems that the strength of discrimination in the direction of the first latent component is greater for assessor "red" than assessor "blue".

3.7 Summary

This chapter concentrated on descriptive analysis and its application to audio evaluation experiments. First, the theoretical background of sensory evaluation practice was shortly discussed. It was concluded that sensory tests can be divided into two approaches with different objectives: affective tests, which are used to evaluate the acceptance or preference of stimuli, and analytic tests, which focus on the distinct perceptual properties of the stimuli.

Furthermore, these two types of testing also require different qualities and skills from the assessors. Typically, naive test subjects are used in affective tests, in which the stimulus is perceived and evaluated as a whole, while experienced assessors, who are able to dismantle their perception into its constituting elements and pinpoint the most prominent characteristics, are used in the analytic tests. Furthermore, the analytic test methods can be categorized into discrimination tests, such as paired comparison and triangle tests, and descriptive analysis methods.

Descriptive analysis was discussed more in detail with verbal elicitation techniques, which are based on the assumption of a tight relation between the perception and its verbal counterpart. These tests have been commonly performed by developing a consensus vocabulary for an expert sensory panel. This is still the most popular approach as it often yields reliable and accurate results, which can be easily interpreted. However, as the time and practical requirements of consensus vocabulary profiling are heavy, and often not feasible, individual vocabulary techniques have been developed to alleviate these issues.

Individual vocabulary techniques were discussed with methods such as The Repertory Grid Technique, Free-Choice Profiling, The Flash Profile, and the most recently developed Individual Vocabulary Profiling by Lorho. Lorho used this methodology to investigate the perceptual properties of sound reproduced over headphones. The results were promising in terms of both the similarity of the perceptual profiles as well as the resource requirements.

Finally, the data analysis of individual vocabulary experiments was discussed. It was concluded that the individual profiles can be investigated for example with PCA and *beta* -coefficient, while the overall combined sensory profiles are obtained with multivariate techniques such as MFA and GPA. In addition, the assessors' performance was discussed with the three most important aspects being repeatability, agreement, and discrimination. In Table 3, an overall comparison of CVP and IVP methods is presented according to Lorho [73].

Table 3: Comparison of consensus and individual vocabulary methods after Lorho [73]

Comparative aspects		Consensus vocabulary	Individual vocabulary
Scope	Sensory charaterization	Quantitative descriptive at a panel level; Validated during the vocabulary development	Quantitative descriptive at an individual level; Validated at a panel level after the vocabulary development
	Stimuli	Relatively flexible in selection and size	Relatively flexible in selection and size
	Project type	High involvement; long-term	Low involvement and short-term
Implementation	Time	low: 15 to 30 hours (or more)	Fast: 2 to 6 hours
	Assessors	Expert permanent panel	Any panel type (consumer panel can be employed)
	Procedure	Panel leader needed; Group work requires careful planning and experience - Improving panel agreement by training is usually a difficult process	Only an introduction to the task and a supervision between vocabulary development steps is needed; No panel leader; Procedure can be semi-automated
	Group work	Needed	Not needed
	Experimental bias	Limited depending on the panel and the panel leader	Minimal
Outcome	Vocabulary characteristics	A single set of sensory descriptors with definition, anchors and sound examplars	A set of individual vocabularies; The large number of attributes brings rich information but limited structure
	Application of the vocabulary	Vocabulary can be reused by the same panel; A new panel can be trained to use the vocabulary	Individual vocabularies can be re-used by the same assessors; Training of other assessors is not possible
	Type of analysis	Relatively simple; Univariate and multivariate analysis	Relatively complex and exploratory; Multivariate analysis only (perceptual directions have to be identified in the latent domain)
	Interpretation of results	Relatively straightforward; Unbiased	Semantic interpretation can be difficult; Biased to some extent

4 A study of perceptual profiling of three finnish concert halls using a Loudspeaker Orchestra and Individual Vocabulary Profiling

4.1 Introduction

As discussed in the second chapter of this thesis, the concert hall acoustics has been traditionally assessed with questionnaires or by comparing recordings or simulations with the attributes defined by the researchers. While the previous work has well established that the listening experience in a concert hall is a multidimensional phenomenon including aspects such as reverberation, loudness, clarity, envelopment, balance and warmth, it is not yet clear which aspects are the most important, particularly in respect of a common concert goer's experience or which are the attributes that would make sense also to the general audience and not only to acousticians and other experts. It was highlighted that the application of attributes which are predefined by the researchers may result in issues of semantic interpretation in evaluation for test subjects.

The methodology of descriptive analysis was discussed throughout the previous chapter. It was stated that traditionally descriptive analysis experiments are carried out with a consensus vocabulary technique in which a panel of assessors develops a common set of attributes for the evaluation of the stimuli. Althought, this approach often yields accurate and reliable results, the time and other practical requirements are sometimes overwhelmingly heavy. Individual vocabulary profiling, on the contrary, applies an individual attribute elicitation and development technique and the comparison and evaluation of the stimuli is performed with these individually produced attributes.

However, one of the main motivations and requirements of IVP is a parallel listening of audio stimuli, which allows *a direct comparison* of various sound features. Clearly, this can not be achieved in natural circumstances in real concert halls. Neither the stimuli can be created by simply recording the playing of a real orchestra in different halls, because the performance is greatly affected by the acoustics of the hall meaning that a real orchestra plays differently in each and every hall. Thus, it is a major challenge in concert hall acoustics research to create such stimuli, which would allow the extraction and evaluation of the very effect of the concert hall acoustics to the listening experience of music. As discussed in Chapter 2, one interesting approach to the creation of the acoustic stimuli is to record the halls using loudspeakers situated on stage playing back anechoic extracts of symphonic music. This method was first used by Göttingen group [93], whose approach consisted of using two omnidirectional loudspeakers on stage and employing an artificial head producing binaural recordings, which were in turn reproduced in an anechoic chamber with two loudspeakers. In many respects, the current study is an advanced and elaborated version of that study conducted over 30 years ago.

In the current study, the stimuli was created by recording concert halls with a controllable virtual orchestra consisting of 34 loudspeakers on the stage in the lay-

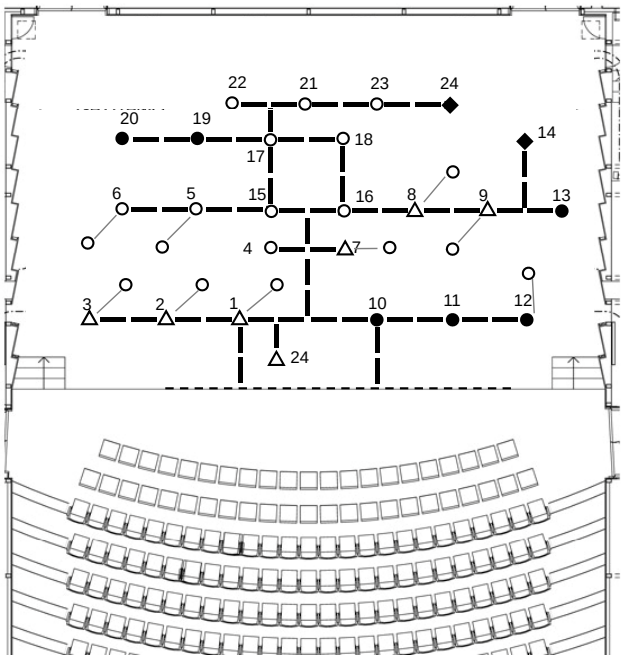


Figure 7: The plan of the loudspeaker orchestra with 34 loudspeakers.

out of a real orchestra (american seating layout [77]) as shown in the Figure 7. The loudspeakers were calibrated and they reproduced the anechoic symphony orchestra recordings of four different musical pieces (compositions by Mozart, Beethoven, Bruckner, and Mahler) exactly the same way in three concert halls [85]. The sound capturing was done in three receiver positions in each hall with a 3-D intensity probe microphone which is shown in the Figure 8. The recorded stimuli were then processed with directional audio coding (DirAC) [100] and reproduced in an anechoic chamber with a 3-D loudspeaker setup consisting of 16 loudspeakers. A picture of the anechoic chamber is shown in Figure 9. This way the variation in the samples was only caused by different acoustics, enabling the parallel listening and direct comparison of the audio stimuli required in the IVP. The signal processing chain is depicted in Figure 10.

A discussion about the making of anechoic recordings, the loudspeaker orchestra, DirAC and the 3-D loudspeaker setup is out of the scope of this thesis, but a detailed description of the equipment and the recording process is reported in Lokki et al. [68]. The focus in this thesis is on the assessor selection procedure and the behavior of individual assessors in the performed IVP study.

This chapter is organized as follows. First, assessor selection procedure is discussed with the emphasis on the triangle discrimination test and the respective results. Then, a detailed description of the IVP experiment is given, with a discussion of prospective modifications to the test procedure. The results of the study are discussed mainly in terms of the individual sensory profiles. The overall results are

Figure 8: A picture of the spatial microphone used in the recording process. The loudspeaker orchestra on the stage is seen on the background.

discussed elsewhere (e.g. Lokki et al. [68]).

4.2 Assessor selection procedure

The assessor selection procedure was inspired by the Generalized Listener Selection procedure [107, 49] originally developed by Zacharov and Mattila. The assessors were selected with a four-phase screening procedure consisting of an online questionnaire, a pure tone audiometry, a test for vocabulary skills, and a triangle test for the discriminative skills of audio stimuli. The screening procedure was developed to meet the requirements of assessing audio stimuli with elicited attributes.

Potential assessors were recruited by sending an email to the students and personnel in the university departments of music, musicology, psychology, acoustics, and media technology. The target population for this study was chosen to be an average Finnish symphony concert audience with some musical background.

4.2.1 Questionnaire

Basic background information, such as a name, age, gender, contact information, and nationality as well as information on musical orientation, interests, hearing, linguistic skills, availability, and motivation for the experiment were collected using an online questionnaire. Two principal pre-selection criteria were normal hearing

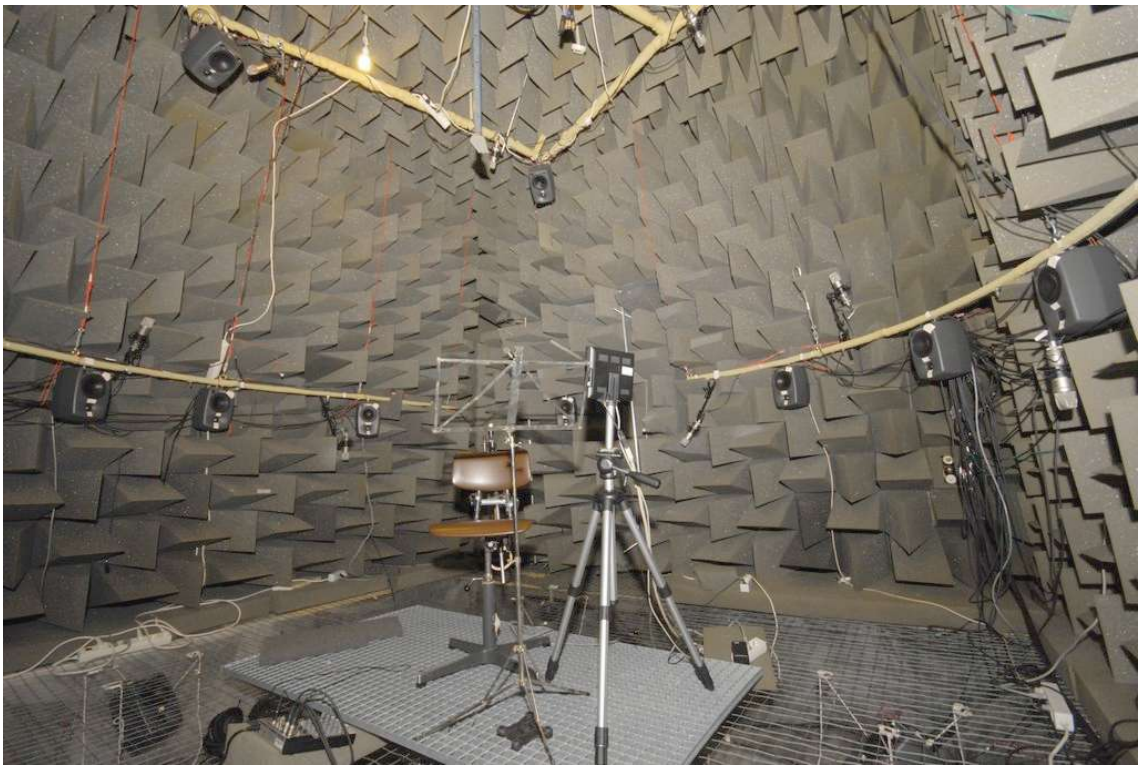


Figure 9: The anechoic chamber, in which the listening tests were carried out.

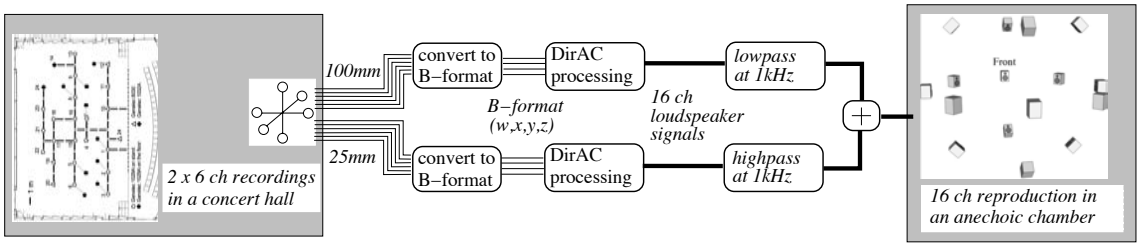


Figure 10: Signal processing chain to obtain comparable stimuli for the subjective evaluation. (25 mm and 100 mm refer to the two spacers used in the spatial microphone probe.)

without any known hearing problems, and native Finnish language, because the individual attribute development process requires good language skills. Availability for testing during working hours and motivation were also important requirements.

A total of 47 people (21 men, 26 women) filled in the online questionnaire and 44 were sent an invitation to participate. Finally 31 candidates responded to the invitation and participated in the rest of the screening procedure.

4.2.2 Audiometry

Pure tone audiometry [50] was applied to check the hearing threshold levels (HTL) of the candidates. The selection criteria were HTLs that should not exceed 15 dB at any frequency band except one which may not exceed 20 dB threshold. Gain steps were quantized to a 5 dB level and an adaptive automated algorithm switched between either a 10 dB decrease or a 5 dB increase in the level of tone depending on whether the candidate heard the tone and responded to the previous trial. The HTL obtained was the level for which the listener responded 3 times out of 5. Completing the audiometry took approximately 15 to 20 minutes depending on the candidate.

4.2.3 Vocabulary test

To assess vocabulary skills and the ability to describe perceptions the candidates were instructed to taste, describe, and compare three orange juices within ten minutes. The use of audio samples was also considered, but it could have been too demanding for the candidates to complete three different listening tests in a row. Thus, it was decided to use three orange juices (Mehukatti-concentrate mixed with water, Rainbow orange juice and Valio orange nectarine) as stimuli and carry out this test between the audiometry and the discrimination test to give candidates a break from listening.

The task was to write down as many descriptive words as possible and to categorize the words according to their modalities. The candidates were instructed to think in terms of different modalities (taste, texture, smell, etc.) and to avoid using adjectives with a hedonic meaning (e.g. "good", "bad", "disgusting", "nice" etc.). On average the candidates produced a total of 15 well defined descriptive attributes. These lists were investigated to verify that the candidates had understood the task correctly and that their vocabulary was adequate for the IVP. Although some candidates clearly performed better than others, no strict selection criteria was applied in this phase. This test also served as a good introduction to the attribute elicitation process.

4.2.4 AAB triangle discrimination test

The objective of the discrimination test was to ensure that the assessors would be able to perceive differences in the audio samples in the IVP. A range of test methods has been developed and used to evaluate the discrimination skills of test subjects (e.g. [11, 65, 49]). Here a forced-choice AAB triangle test with replications was employed. The candidates were presented a set of sample triads and the task was to detect the sample that differed from others in each triad. The theoretical considerations of this method are well discussed, e.g., in [19, 82, 57].

The audio material to be used in a discrimination test has great importance as it dictates the perceptual differences that are detectable between the stimuli. There have been several different approaches to the selection of the sound samples. Most often the discrimination skills have been tested with a simple loudness test using pink noise [65]. In addition, speech and audio quality tests use varying encoding methods

in order to produce the desired perceptual differences [49]. Furthermore, Isherwood et al. [49] have elaborated the discrimination tests to include the perceptual testing of spatial differences in loudspeaker and headphone reproduction. In another study, Lorho [71] selected the samples to include also differences in timbral as well as in spatial aspects.

To correspond with the particular setting of this study, it was desirable to use the newly recorded material also in this phase. Additionally, the discrimination test was considered providing some preliminary information on the recordings. Thus, the samples for the triangle test were selected and extracted from binaural recordings (B&K 4100 sound quality HATS) of the loudspeaker orchestra playing back a part of an aria of Donna Elvira by W. A. Mozart. The recording positions corresponding to the samples are illustrated in Fig. 11.

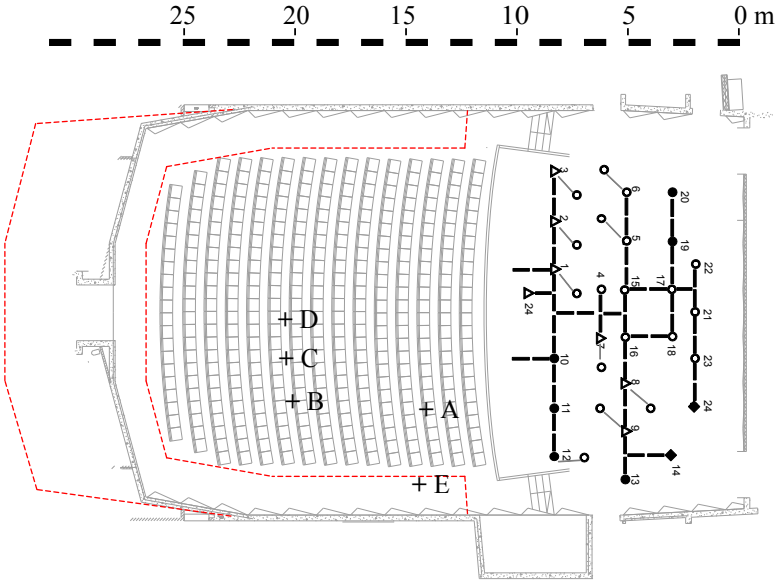


Figure 11: Recording positions from which the samples were selected for the discrimination test. (Position "E" is on the balcony.)

Table 4: The sample pairs in the discrimination test

Pair	Samples	
1	A	E
2	A	B
3	B	D
4	B	C

Four sample-pairs shown in Table 4 were selected to represent varying levels of perceptual differences. The degree of the difficulty of the task was considered changing in the respective order of the sample pairs, that is, pair 1 being the easiest

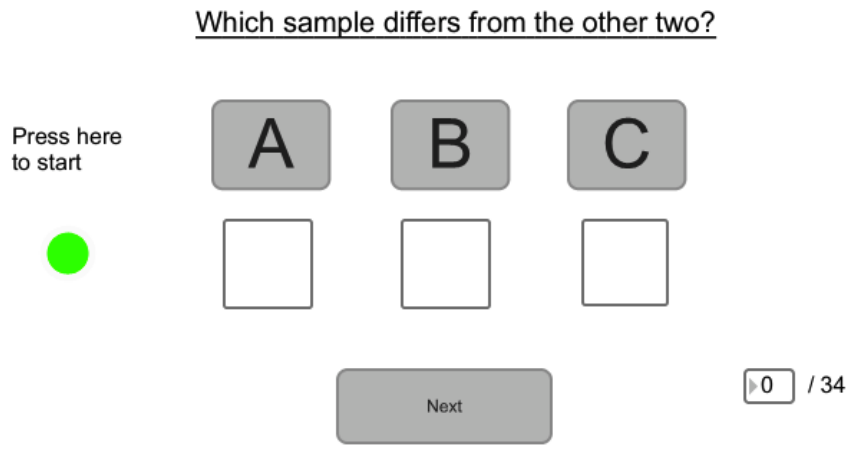


Figure 12: The graphical user interface used in the triangle discrimination test.

and pair 4 the hardest pair to judge. However, this assumption was revealed to be deficient in the analysis of the discrimination test data.

The triangle test was implemented with MAX/MSP 5 graphical programming language/environment. The test was designed according to the descriptions in [65] and [71] with minor modifications. The test started with an introductory sequence of the easiest sample-pair and these triads were presented in the following balanced sample order: ABB, AAB, ABA, BAA, BBA, BAB. This introduction ensured that the candidates had understood the task correctly and could operate the GUI appropriately. The data from this learning sequence were not included in the analysis.

Next, the presentation order of the sample-pairs and the triads were randomized in the following way. For each sample-pair, each of six triads described earlier were first presented once in a random order, and then a random supplementary triad was selected from the six triads. Thus, there were seven triads for each of the four sample-pairs constituting a total of 28 triads for the whole test. Furthermore, the presentation order of these 28 triads was randomized.

The crossfade switching time between samples within a triad was 750 ms with 200 ms linear fade-in period. The samples started from the beginning when the corresponding button was clicked on the GUI. The candidates had to listen to each of three alternatives at least once and to give an answer before it was possible to move on to the next triad (even if they could not identify the odd sample). Additionally they were able to choose only one alternative for an answer. These aspects were programmed to avoid unintentional button presses and other mistakes as well as to simplify the use of the GUI. A picture of the GUI is presented in Fig. 12.

By implementing the original test program, it was possible to decide exactly what kind of information was collected. As discussed also by Legarth and Zacharov [65], it is desirable to not only get the indications of the correct answers, but also some information on the behavior and performance of the candidates. In this study, additional information included the number of switching between samples for each

triad, the response time for each triad and the time spend for completing the test. However, the analysis of the supplementary data is outside the scope of this thesis.

4.2.5 Data analysis of triangle test

The analysis methods for replicated triagle tests have been elaborated and discussed by Brockhoff and Schlich [23], Kunert and Meyners [57] and more recently by Brockhoff [24], Duineveld and Meyners [33], Bayarri et al. [10] and Meyners and Duineveld [78]. In this study, the analysis of the triangle test data was twofold. First part of the analysis consisted of investigating if the candidates were able to discriminate the samples in a statistically significant way. Secondly, the results were used for investigating the properties of the constructed sample pairs by conducting a simple test of proportions of correct answers.

Here, a total of seven triads for each sample pair were presented. In a triangular forced-choice test the probability for guessing the odd sample in a single triad is $p_0 = 1/3$. Consequently, the minimum number of correct answers to establish significance in a triangle test is obtained from a cumulative binomial probability distribution. The values are 7 on the significance level p=0.001, 6 (p=0.01), and 5 (p=0.05). The scores from the triangle test for each candidate and each sample pair are shown in Table 5. A limit of 5 correct answers to the pairs 1 and 2 as well as at least one of the pairs 3 and 4 was employed as a selection criteria.

In the second part of the analysis, it was investigated if significant differences existed between the pairs with a Z-test for two proportions. The total numbers of corrects answers to each sample pair are shown in Table 5 and the results of the proportions test is shown in Table 6.

The null hypothesis for this test was that two proportions are not significantly different (H_0: $p_1 = p_2 = p$) and the alternative hypothesis was one-tailed, so that the easier pair would have a proportion larger than the harder one (H_1: $p_1 > p_2$). The test was done between the pairs 1 and 2, pairs 2 and 3, as well as 3 and 4. Note that the test results show that considering the pairs 1 and 2, we would reject the null hypothesis at a confidence level of 95 % ($p < 0.05$) but not at the level of 99 % ($p > 0.01$). In addition, the results show that considering the two hardest pairs, the proportions were not significantly different. In other words, the degree of the perceptual differences between places B and D and between places B and C were not significantly different according to these results.

The implications of these results are important: First, the validity of the triangle test with original audio samples should always be verified before application to a screening procedure. In this case, it would be necessary to construct new sample pairs which would represent a wider range of perceptual differences. Second, if the degree of perceptual difference between places B and C does not significantly differ from that between places B and D, there might be no reason to make recordings in all of these places. However, one cannot conclude that there would be no differences at all.

Considering that only a few IVP studies exist in the whole audio field and none concerning the concert hall acoustics, it was finally desired to have as many assessors

Table 5: Results of the triangle test.

	Number of correct answers				
Subject #	pair 1	pair 2	pair 3	pair 4	total
1	7	7	**4**	5	23
2	7	7	6	**2**	22
3	7	7	5	5	24
4	7	**4**	**3**	**4**	18
5	7	7	6	**4**	24
6	7	7	**2**	**1**	17
7	7	7	**4**	**4**	22
8	7	7	5	**4**	23
9	7	7	**4**	**2**	20
10	6	6	6	**4**	22
11	7	7	**3**	6	23
12	7	7	5	7	26
13	7	7	5	**3**	22
14	7	7	5	**4**	23
15	7	6	**3**	**2**	18
16	6	5	**3**	**4**	18
17	6	6	**3**	**3**	18
18	7	7	**4**	**4**	22
19	7	7	**4**	6	24
20	6	6	**4**	7	23
21	7	7	5	**3**	22
22	7	**4**	**1**	**3**	15
23	7	7	**3**	**4**	21
24	7	7	**4**	**4**	22
25	7	7	**3**	**3**	20
26	7	7	5	7	26
27	7	7	**1**	**3**	18
28	7	6	6	**4**	23
29	7	7	5	**4**	23
30	7	7	6	**3**	23
31	7	7	7	6	27
Total	213	204	130	125	672/868

Table 6: Proportion test results (one-tailed).

Test pairs	Z-value	p-value
1 and 2	1.981	0.024
2 and 3	8.312	<0.001
3 and 4	0.391	>0.1

as possible for the listening tests. Thus, 20 candidates were selected not only on the grounds of the audiometry and the discrimination test but also on the grounds of motivation and availability. Eighteen candidates passed the audiometry and the discrimination test. Two additional candidates were still selected as assessors as they were highly enthusiastic and motivated. All selected assessors were Finnish university students of music, musicology, psychology or acoustics.

4.3 Implementation of the Individual Vocabulary Profiling Procedure

The individual vocabulary profiling consists of elicitation and development of individual attributes for a comparative evaluation of the stimuli. In this study, each listening session lasted a maximum of two hours and the whole sensory profiling was completed in four sessions per assessor. All sessions were held in an anechoic chamber with a dim lighting. The sound reproduction system had 16 loudspeakers in a 3D setup, see Figure 9. Each assessor completed the procedure individually. This resulted in a total of 160 hours of listening tests. Assessors were also interviewed after the experiment and asked to fill a small inquiry of their general impressions. They were also asked if they had used any systematic strategy in the evaluation process. The following sections describe the whole procedure in detail.

4.3.1 First session

The main objective of the first listening session was to serve as an introduction to the test procedure. It consisted of a familiarization with the sound material and the graphical user interface (GUI) as well as elicitation of a preliminary list of descriptive words.

First, assessors were presented with the general aspects of the study verbally and in writing to give an overview of the research. It was also verified that they were comfortable to do the listening tests in the dim anechoic room as this could cause anxiety in some people. The assessors were also told about possible artifacts during the playback (e.g. background noise, clicks and pops etc.) that they were not supposed to pay attention for. Then they listened to the whole sample set (nine samples of each of four different musical compositions) without any specific task. Then they were instructed to listen to the samples using the GUI, depicted in Fig. 13. At the same time, they were asked to search for any perceptually interesting aspects of the sounds to ensure a proper familiarization to both the audio material and the user interface. The GUI enabled the looping of a selection for listening to a particular part if needed.

During a short break, the assessors were presented a pre-made list of sound related descriptive words, which have been previously used and developed in the audio field. This was considered facilitating the upcoming first elicitation phase, but it was also stressed that this list was only for a reference and their own terms and attributes did not have to be from this list. After the break, assessors continued to listen to the samples but now they were instructed to write down all descriptive

words that came into their mind. The whole sample set was played back with no requirements to use the interface as it could have only mixed up the free elicitation of words. This phase ended the first session.

4.3.2 Second session: The development of the attributes

At the beginning of the second session, assessors listened to the whole sample set and reviewed their own preliminary attribute lists made in the first session. They were free to add or discard words if they found several words describing the same aspect in the sound samples. In addition, they were instructed to select the most appropriate and descriptive words in their list and to eliminate words with hedonic and affective connotations. The goal was to condense into the 4 to 6 most descriptive attributes which could be used to discriminate these audio samples.

Then, the assessors were asked to review their list of 4-6 attributes and to write down brief descriptions of the respective perceptual aspects. They also defined the respective bipolar anchor labels for the continuous scales. This was carried out under the supervision of the experimenter to ensure that the developed attributes were descriptive without affective connotations although the experimenter gave as little advice as possible to prevent any bias.

The attribute development process usually took two hours. If there was enough time, a first practice session with the whole stimulus set was held right after forming the attributes. This helped the assessors to have an immediate impression of the suitability of their attributes and how their anchor labels worked in the evaluation. At this point, it was advised that the assessors would start using the scales and search for the samples that represented the extremes of the given attribute.

4.3.3 Third session: Dress rehearsal

The third session consisted of a simulation of the sensory profiling task. The assessors completed the assessment with their own attributes and definitions. Additionally, they were instructed to recheck their attribute list once more and after completing the task, they were able to make final modifications, or even add or remove attributes. Most often only minor adjustments, if any, were seen necessary.

4.3.4 Final session: The real thing

In the final session, the assessors completed the sensory profiling task with their own attributes. They also had access to their own definitions during the listening. The presentation order of nine samples in each window (corresponding to a particular attribute - composition pair) was fully randomized as well as the presentation order of attribute - composition pairs. For example, if the assessor had developed 5 attributes, the whole evaluation consisted of 20 (5 attributes times 4 musical excerpts) sets of 9 samples. The assessors were strongly advised to have at least one short break to maintain an adequate concentration and performance level.

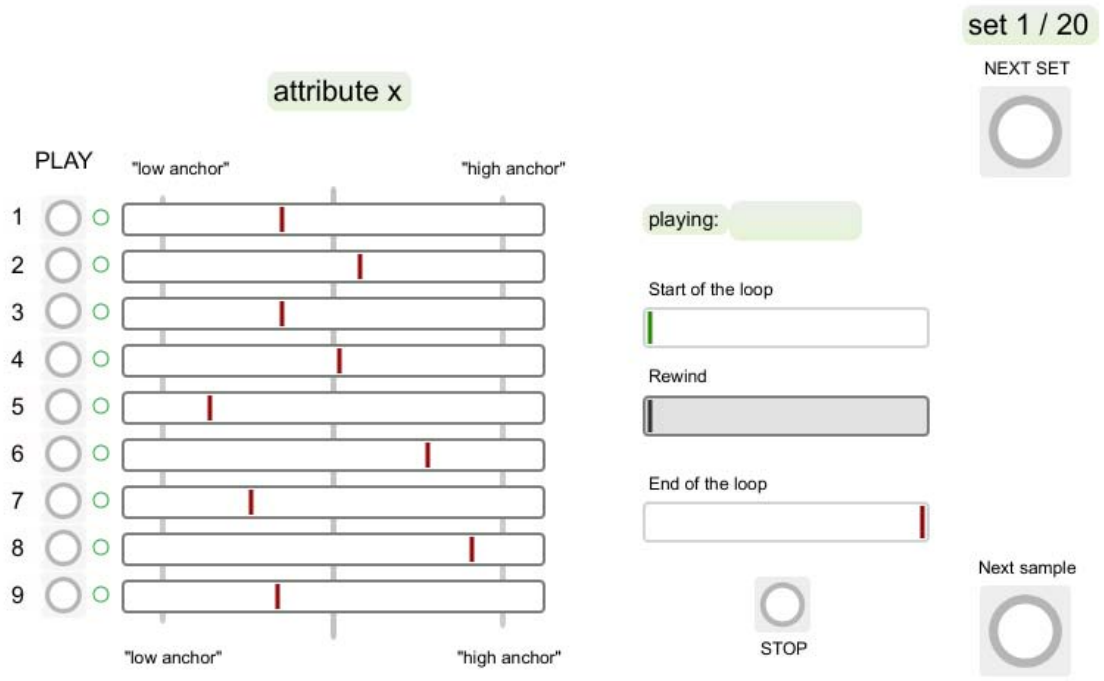


Figure 13: The GUI for the individual vocabulary profiling. (Translated into English.)

4.3.5 Notes on the experimental design

The experimental design of this sensory analysis experiment was adapted from the Flash Profile method developed by Dairou and Sieffermann [29] and the IVP developed by Lorho [73]. The modifications were made according to the requirements and resources of our study. The following notes on the further developments of this test procedure were made during the listening tests.

A verbal direct attribute elicitation technique might not be the most effective in the case of naive assessors. Alternatives such as Repertory Grid Technique [55] could result in an unequal set of attributes. RGT has been previously applied in the spatial attribute elicitation by Berg and Rumsey [17] with interesting results. Generally, attribute elicitation techniques have been of great interest in the field of consumer research and some comparisons between different techniques can be found, e.g., in [22, 12]. However, considering that one motivation of the IVP is its short time requirement, elicitation methods such as RGT might not come in question. Comparison of different verbal elicitation techniques in the audio field would be needed to determine the suitable alternatives to the direct attribute elicitation technique in the IVP.

The scales in this experiment were continuous, but as in the Flash Profile method [29], the use of nominal scales could be beneficial facilitating the evaluation process. On the other hand, if it was only desired to get the order of the samples, some

kind of a sorting algorithm such as Quicksort with the assessor as the decision maker could be developed. Presumably, pair-wise comparison would be lot easier than the comparison of several samples at the same time. This would also remove the need to divide the samples into groups in terms of musical piece or any other feature. However, the information that would be lost in this approach is the distances between the samples what is obtained by using the continuous scales and parallel comparison. Nevertheless, it would be also interesting to see, if the consistency of and the agreement between assessors could be augmented this way.

Finally, the implementation of original test programs allows the collection of various information on top of the evaluation results. For example, considering the reported evaluation strategies in this experiment, it would be interesting to investigate the attentional focal points in the samples. It would be valuable to know what kind of audio events are being listened to in the evaluation of a particular attribute. This knowledge could then be used in the production of stimuli in the future studies of concert hall acoustics.

4.4 Analysis of individual sensory profiles with PCA and *beta* -coefficient.

There are many ways to analyze the data from the individual vocabulary profiling experiment as discussed in the previous chapter. Here, the focus is on the individual behavior of assessors by investigating the perceptual spaces with principal component analysis (PCA) and *beta*- coefficients. Although, the analysis of variances (ANOVA) and associated models would be also feasible, these are not presented in this thesis for two reasons: firstly, the abundance of data from the experiment, i.e., 102 attributes which should all be evaluated separately, and secondly, ANOVA has already been extensively used in the literature and author's inclination is to present a less familiar analysis approach with the *beta*- coefficient. The data analysis was performed with the FactoMineR package [63] and Matlab. A full analysis of the combined data of all assessors is beyond the scope of this thesis and it is presented in [68].

4.4.1 Individual sensory profiles with PCA

PCA calculates the uncorrelated principal dimensions (components) corresponding to the variances in the data. This kind of analysis does not offer any definitive conclusions and has to be carefully interpreted in order to not to make false assumptions. For example, if two attributes are highly correlated in the perceptual space, one simply can not conclude that the assessor has evaluated these attributes using the same perceptual criteria. The perceptual aspects behind these attributes might be just as well correlated and varying in the same direction (which is often true).

However, some information on the individual behavior can be extracted with PCA. By investigating the coefficients of determination as well as the correlations between different attributes, one can evaluate the complexity of the individual per-

ceptual spaces. For example, if there are many attributes with high correlations, it is possible that these attributes carry a lot of redundant information. If the attributes are well spread across the whole perceptual space, they arguably contain information on various aspects of the stimuli. It is a question of how many perceptual directions can be easily identified in the latent sensory space spanned by the principal components.

To illustrate this way of thinking, let's consider the sensory configurations of two assessors. In this example, all the data of each assessor is included in the analysis, that is, the matrices of 36 rows (9 samples per each of four musical pieces) by the respective number of attributes were created. For these matrices, principal component analysis was performed in the R environment. Representations of the PCA results are plotted in Figure 14. These graphs illustrate the attribute correlations and the sample positions in two principal dimensions as well as the respective confidence (95 %) ellipses corresponding to different halls and musical pieces. For the other 18 assessors, the similar attribute correlation graphs are shown in the appendix in Figures A1, A2 and A3. Also, all elicited attributes with the anchor labels are presented in the appendix in Table A1.

First, the differences between individuals can be evaluated by looking into the explained variation by the first two components. Concerning assessor 9, the first principal dimension already explains 81.7 % of the whole variation in the data and the second only 6.5 %. Additionally, in this case all of the attributes are highly correlated with the first dimension and only two separate perceptual directions can be identified. As stated previously, this kind of unidimensionality indicates that this assessor may not have been able to discriminate between the attributes or there has been some governing aspect which has strongly influenced the perception. This could be also due to the inexperience of this subject and further training could be beneficial as stated by Labbe et al. [59].

For the assessor 15, the first axis explains about 60 % of the variation and the second also 26 %. It could be even reasonable to look into the third dimension. The attributes are more spread around the circle and it seems that each of them defines a proper perceptual direction. This clearly indicates that different attributes have been clearly related to different kinds of perceptual features and the vocabulary development process has been succesful. The middle graphs in Figure 14 show the sample positions in the first two dimensions and they also reflect the attribute correlations: for assessor 9 the samples are condensed around the first axis while they are more spread around for assessor 15. Thus the interpretation of the results for assessor 15 is much easier than for assessor 9.

For example, concerning assessor 15, the sample group of four samples on the left-up side of the graph (abbreviated with kor6) have been perceived to have a high level of drr and envelopment, while most of the samples from Tapiola hall (abbreviated with "tar") on the right-down side have a high level of drr but have a low envelopment and eq. Considering assessor 9, interpretation of the results is more one-sided, the samples differ greatly only in the first dimension which explains most of the total variance. For example, it can only be concluded that the samples which have perceived to be loud (high volume) have also been perceived to be clear, deep,

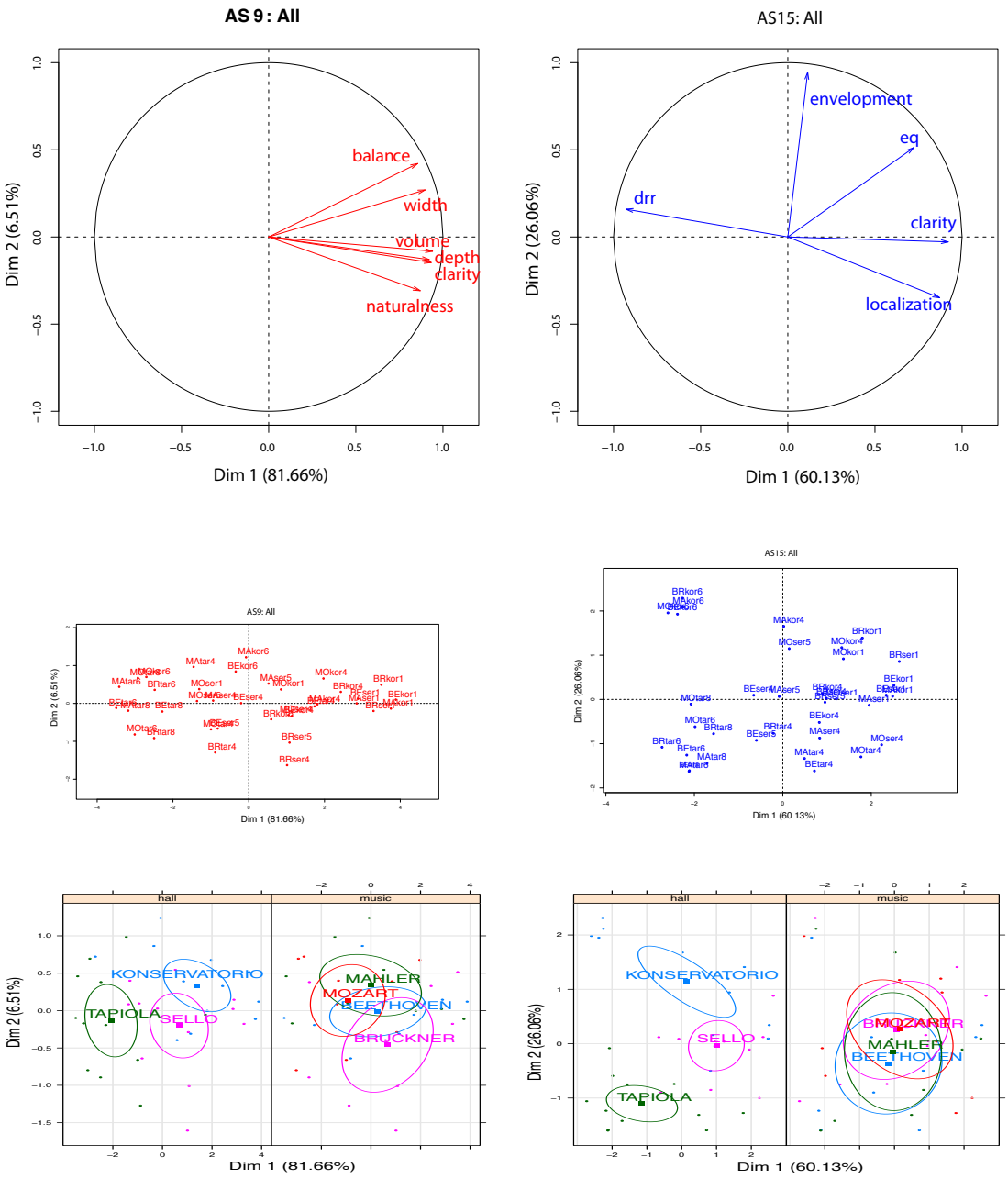


Figure 14: An example of comparison of the individual sensory profiles of two assessors with PCA. The abbreviations in the middle graphs are MO = mozart, MA = Mahler, BE = Beethoven and BR = Bruckner for signals, and ko = Konservatorio, ta = Tapiola and se = Sello for halls. For example BRkor6 means position 6 in Konservatorio hall with stimulus signal Bruckner.

natural, wide, and balanced. As the same conclusion can be drawn with all attributes and samples, it is clear that this information is rather one-sided; there seems to be only one clear perceptual dimension described with all of these attributes. Note,

that these aspects also relate to the concept of discrimination, which is one of the main aspects of the performance consideration of assessors. The terms strength and the dimensionality of discrimination defined by Shlich et al. [91] can be used in this respect. The correlation graphs combined with the product spaces show that while the profile of assessor 15 show the greater dimensionality of discrimination, the strenght of discrimination is greater for assessor 9 regarding the first principal component.

The third graphs show the sample groups, which have been significantly different from each other. It can be clearly seen that, in overall, the hall has influenced the perception more than the respective musical piece. To be specific, the samples from the Tapiola concert hall have been the most distinguishable. The musical piece has influenced only a little the perception and evaluation of the samples with these attributes. Still it can be noted, that the musical piece has influenced more the perception of assessor 9 than assessor 15.

Table 7 contains the main perceptual directions interpreted in the latent sensory space spanned by the first two PCs of the associated PCAs. In overall, it can be clearly noted that there is generally one perceptual direction related to distance and loudness, a second one related to reverberation, a third one related to width, and a fourth one related to timbral aspects such as clarity, definition and brightness. However, often the timbral aspects are correlated also to loudness and distance attributes indicating some confusion in this respect. It is also interesting to note that there are often two groups of spatial perceptual directions (exluding distance): reverberation and width or envelopment.

The grouping of individual attributes based on the attribute rating data can be performed with Hierarchical Cluster Analysis and it offers a more analytical approach to evaluation of the agreement of assessors in terms of different attribute groups. However, this analysis is out of the scope of this thesis. It is presented with other overall results in [69], and [68]. On a general note, these results are in accordance with the interpretation presented here.

4.4.2 *beta*- coefficient as a measure of complexity of the individual sensory profiles

As discussed in the previous chapter β-coefficient can be used to evaluate the dimensionality or complexity of a data matrix. Here, it is used as a measure of the complexity of the individual sensory profiles and the results are compared with the PCA solutions. The variances explained by the uncorrelated principal components (dimensions) can also be considered as indicators of the dimensionality of a sensory configuration as discussed previously. However, a classic problem with PCA is to determine the criteria which is used to decide the number of components included in the interpretation of results. Common approaches are to include the components which explain more than 10 percent of the total variation in the data (used in this work) or the components of which the associated eigenvalue is more than 1. Also, one may determine the "knee" point where the explanatory power of components vanishes by plotting the associated eigenvalues in a simple graph.

Table 7: The main perceptual directions interpreted in the latent sensory space spanned by the first two principal components of the associated PCAs (see Figures A1 and A2 in Appendix).

AS	Main perceptual directions
AS1	(1) width, (2) loudness/separation, (3) clarity,
AS2	(1) distance/width, (2) reverberation, (3) transparency
AS3	(1) distance/intimacy/approach of sound, (2) width, (3) clearness
AS4	(1) naturalness/sense of space/stand out/full-flavored, (2) symmetry
AS5	(1) loudness, (2) reverberance, (3) emphasis on bass, (4) closeness/balance
AS6	(1) loudness/brightness/closeness/liveliness, (2) reverberance
AS7	(1) distance/envelopment/full flavored, (2) reverberation, (3) openness/definition
AS8	(1) closeness/clarity/dynamics, (2) broadness, (3) tone color/definition
AS9	(1) volume/depth/clarity/naturalness, (2) width/balance
AS10	(1) loudness/distance/muddy/amount of bass, (2) wideness, (3) amount of reverb
AS11	(1) soulless/naturalness, (2) precise, (3) wide
AS12	(1) clearness/distance, (2) definition, (3) balance
AS13	(1) distance/volume, (2) reverberation, (3) directed
AS14	(1) closeness/spread of sound/texture, (2) clearness, (3) 3-dimensional
AS15	(1) clarity, (2) eq, (3) envelopment, (4) drr, (5) localization
AS16	(1) distance/loudness, (2) softness, (3) width, (4) reverberance
AS17	(1) distance of sound source, (2) brightness, (3) tone color, (4) reverb
AS18	(1) distant/neutral/presence, (2) wideness, (3) reverberant, (4) pronounced
AS19	(1) localizability, (2) distance/sharpness, (3) definition, (4) size of space/sonority
AS20	(1) bass/focused sound, (2) distance/treble, (3) balanced, (4) reverb

The β-coefficient however is a single measure of the dimensionality of a data matrice, and offers an alternative approach to the complexity evaluation. It can be used as a simple indicator of the number of dimensions or ideal attributes sufficient to completely describe a set of stimuli. But, as Schlich [90] discusses, it should not be thought as an indicator of the number of attributes to be elicited and used in the evaluation. The number of attributes should be at least the double of the calculated β-coefficient. The β-coefficients per subject and musical piece are tabulated in Table 8. A Matlab routine was implement for the calculation and it is included in Appendix. Also the number of PCA components explaining more than 10 percent of the total variation in the data are shown here for comparison.

Let's first consider only the β-coefficients. The results indicate that the number of perceptual dimensions which can be used to describe these stimuli and to interpret the results is from 2 to 3. It is also interesting to note that the complexity of the individual sensory profiles seems to be independent of the number of attributes used in the evaluation. This suggests that the complexity of the profile depends on the abilities of the assessor rather than the number of attributes. For example, assessor 3 has developed 5 attributes but the complexity remains low in each case (1.2 - 1.6), while the *beta* -coefficients for assessor 1 are greater (1.8 - 2.9) although he/she has used only 4 attributes in the evaluation.

Table 8: Complexity of the sensory profiles described with β-coefficients and the number of PCA components explaining more than 10 percent of the total variance.

	#	All		Moz		Bee		Bru		Mah	
AS	attr.	β	PCs	β	PCs	β	PCs	β	PCs	β	PCs
AS1	4	2.5	3	2.9	3	2.1	2	1.8	2	2.1	2
AS2	4	2.1	2	2.1	2	1.8	2	1.6	2	1.6	2
AS3	5	1.4	1	1.3	1	1.6	2	1.2	1	1.2	1
AS4	5	2.3	2	2.6	3	2.1	2	2.0	2	1.6	2
AS5	5	3.4	3	2.0	2	2.2	2	2.3	3	3.0	3
AS6	5	1.7	2	2.0	2	1.7	3	1.6	2	1.4	1
AS7	6	1.9	2	2.3	3	1.7	2	1.7	2	1.6	2
AS8	6	2.4	3	1.9	3	2.4	3	2.6	3	2.0	2
AS9	6	1.5	1	1.8	1	1.5	1	1.6	1	1.3	1
AS10	6	1.9	2	2.2	3	1.3	1	1.9	2	1.5	2
AS11	4	2.3	3	2.8	3	1.8	2	2.1	2	1.3	1
AS12	4	2.5	3	2.2	2	2.3	3	2.6	3	2.1	3
AS13	4	2.3	3	2.5	3	1.6	2	2.0	2	1.9	2
AS14	5	2.4	2	2.6	3	1.6	2	2.2	3	2.4	3
AS15	5	2.3	2	1.9	2	2.2	2	2.1	2	2.5	2
AS16	5	2.7	2	1.6	2	2.6	3	2.2	2	1.9	2
AS17	5	2.8	3	2.1	2	2.2	3	2.5	2	2.8	3
AS18	6	2.1	2	2.0	2	1.7	2	2.0	2	1.7	2
AS19	6	1.8	2	1.3	2	1.7	2	2.3	2	2.0	2
AS20	6	2.9	2	3.0	3	2.4	2	2.1	2	2.5	3
MEAN	5.1	2.3	2.3	2.2	2.4	1.9	2.1	2.0	2.1	1.9	2.1

Regarding the different musical extracts, Mozart has the greatest average complexity while Beethoven and Mahler share the lowest, although the differences are small across the different pieces. Nevertheless, it may be speculated that the different acoustical cues have been clearer in the other pieces, while in other's they are masked, e.g., by the powerful style typical to Mahler's symphonies resulting in the reduced complexity of the sensory profiles. Moreover, some assessors have a quite large variation of the perceptual complexity between the musical extracts what reinforces the common assumption that the type of music influences the perception of the acoustical properties of the halls depending on the individual. The largest and smallest differences are shown for assessor 11 (2.8 - 1.3) and assessor 18 (2 -1.7). This indicates that some assessors seem to be able to identify the acoustical aspects despite the varying musical styles, what may be regarded to reflect a certain sensory skill that these assessors possess. This is typically also shown in the PCA results represented by attribute correlation graphs (see Appendix).

By comparing the β-coefficients and the number of PCA components, it can be clearly noted that they are essentially similar, which arguably proves the validity of

the β-coefficient as a measure of the complexity of a sensory profile. The β-coefficient being more accurate, it offers a more analytical way of comparing the complexities of the individual profiles although it does not offer any additional information about the perceptual characteristics of the stimuli, what is the case with PCA. On a final note, β-coefficient is a good complement to PCA as it reduces the speculation about the number of dimensions to be included in the interpretation of the results.

4.4.3 Interviews

After the completed test, the assessors were asked to fill in an enquiry concerning the practical issues of the test procedure. In addition, they were asked if they had used any conscious strategy in the evaluation and if they had some additional comments on the test procedure as a whole.

In short, the degree of difficulty of the listening tests was given an average score of 3.7 (on the scale 1 "easy' ' – 5 "hard") and the assessors rated the procedure as highly interesting with an average score of 4.5. They also considered the GUI easy to use with a score of 1.5 (1 representing "really simple") as well as the time reserved for the completing listening tests to be adequate with a score of 2.9 (1 representing "not enough" and 5 "too much"). Finally, they rated the audio quality of the samples with an average score of 3.6 out of 5. A possible reason for this lower score of audio quality is speculated to be the low-level noise present in the samples of musical extracts of Mozart.

The reported strategies most often consisted of searching a short loop in the samples representing the particular attribute and then comparing the samples using only this selected period. Comparison was most often started with determining the samples which represented the extremes of the scale and then relating the rest of the samples to these extremes. The intermediate comparison was usually done with a pair-wise comparison strategy. Some assessors also reported on difficulties in the evaluation of some compositions with some particular attributes and suggestions were made to a development of attributes which would be composition specific.

4.5 Discussion and conclusions

4.5.1 Screening procedure

The screening procedure consisted of an online questionnaire, a pure tone audiometry, a vocabulary test and a triangle sound discrimination test. From 47 candidates, who responded to the questionnaire, 31 participated in the rest of the screening and finally, 20 assessors were selected for the experiment according to their abilities and motivation. These numbers alone point to the importance of having a large initial group of potential test subjects, as typically there is a large percentage of people, who do not meet the requirements or are not available or motivated enough to finish the listening tests, as described also by Bech and Zacharov [11]. Furthermore, although the selected assessors all had some sort of musical background, they were all more or less inexperienced in terms of sensory evaluation and descriptive analysis.

The selection procedure was developed to meet the requirements of the individual vocabulary development process. After the audiometry, a vocabulary test with orange juice stimuli was performed before the sound discrimination test. This kind of vocabulary test was inspired by Legarth and Zacharov [65], who applied a similar approach in the selection of assessors to multisensory applications. It was also thought to give some time to the candidate's ears between the audiometry and the discrimination test.

However, in author's view this was not a good choice as it turned out to be impossible to apply any screening criteria to the vocabulary test results. Also, it would have been more justified to use audio stimuli in this phase, as it is clear that describing tastes is different than describing sounds. Even a better alternative would have been to integrate the vocabulary elicitation phase into the triangle discrimination test, what would have resulted in a similar approach, which is applied in the Repertory Grid Technique. This method has been also used by Berg and Rumsey [18] and Choisel and Wickelmaier [28]. Moreover, this way the triangle discrimination test could have served also as the first attribute elicitation phase in the IVP further reducing the time requirement of the experiment and make the first elicitation more formal and structured.

4.5.2 Individual vocabulary profiling experiment

The naive assessors quantitatively evaluated 36 stimuli - 9 different acoustics and 4 different musical extracts - with each of their own attributes. The stimuli were recorded by using a controllable virtual orchestra consisting of 34 loudspeakers to allow a simultaneous comparison. Here, the focus is on the individual sensory profiles obtained with PCA and the complexity estimation of the profiles with the *beta* -coefficient. The results indicate clear differences in the assessors' abilities to produce attributes, which would discriminate the stimuli in many perceptual aspects. The PCA showed for some assessors very high correlations between several attributes what arguably points to some difficulties in the attribute development process. Other assessors were clearly better breaking their perception into its constituting elements, what was expressed by an easy identification of the salient perceptual directions in the latent sensory space. In overall, the most distinguishable perceptual aspects were related to loudness, distance, reverberation, width and definition/clarity.

A complementary view to these results was given with the *beta*-coefficients, which indicate the number of ideal attributes sufficient to fully describe this set of stimuli. This complexity measure was on average between two and three, but differences between assessors were manifested in this respect. These differences were generally in accordance with the PCA results, i.e. the sensory profiles with easily identifiable perceptual directions also had greater complexity. In the field of audio, also Lorho [73] has used the *beta* -coefficient as a complexity measure of the individual profiles of an IVP study of sound reproduction over headphones. In his study, the complexity measures were a little higher, from 2 to 4, but also the number of attributes were greater, from 4 to 8. In both studies, the vocabulary size did not automatically result

in a greater complexity and there was a decline of complexity across different musical extracts. This raises a question about the musical properties resulting in difficulties of perceiving the characteristics under evaluation (arguably manifested in a reduced complexity of a sensory profile). Futhermore, in the field of acoustics, could it be possible to use the acoustical properties of a hall in relation to the musical style to compensate this perceptual change? The reasons for this are however still unclear.

To evaluate the reliability and consistency of the assessors, some form of repeated measurements would have been needed. The applied methodology described previously unfortunately did not comprise such repeated tests per se, as the assessors were given the possibility to adjust their attribute list between the practise run and the final evaluation. These last minute adjustments usually included removal or change of one or two attributes and modifications on the anchor labels and although, they were often seemingly minor, they were however, in author's view, badly needed. Nevertheless, these adjustments prevented a straightforward analysis of the performance of the assessors, even though the data from the practise runs were collected and saved. The importance of careful planning in performing experiments can be clearly noted in this respect.

It is clear that assessors' repeatability can not be evaluated without repeated measurements. However, in the framework of IVP, agreement between assessors can be evaluated with several approaches. One way is to group individual attributes qualitatively or quantitatively into separate perceptual categories. Qualitative grouping can be done by looking into the attributes, anchor labels and descriptions and this way the experimenter can take into consideration the semantic meanings of attributes. However, this kind of analysis may be easily biased by the experimenter's interpretation as discussed by Lorho [73].

The grouping of attributes can be also performed quantitatively on the basis of attribute rating data. One common method is Hierarchical Cluster Analysis (not presented in this thesis), which basically calculates the Euclidean distances between attributes and categorizes them into clusters accordingly. This is also a good way to analyse different scenarios present in the vocabulary of assessors (i.e. consensus, correspondence, conflict and contrast) Considering the current study, a detailed analysis of the attributes and their clustering is presented in [69] and [68]. In brief, the attributes form 8 groups including two for reverberance (size of space and enveloping reverberance), apparent source width, balance, loudness, distance, openness, and definition. In addition, eleven attributes can not be grouped based on their definitions, but all of them correlate highly with loudness and distance.

Another approach to assessing the agreement between assessors is to derive a consensus sensory profile with multivariate techniques such as (H)MFA and GPA. For example, it is discussed in [68] how the behaviour of assessors in relation to the consensus sensory profile can be evaluated by MFA. It was shown that the consensus was well established in the first principal dimension (loudness/distance/openness/ungrouped) of the MFA solution, while the assessors performed more differently regarding the second (reverberance) and other dimensions.

Still, one, a little more experimental approach to the evaluation of agreement between profiles, would be given with the RV coefficient, which is the generalized

Pearson correlation coefficient between two data matrices. Theoretically, the RV coefficient would indicate the level of similarity between two sensory profiles in terms of the structure of the perceptual space. However, this is somewhat problematic in the framework of IVP, in which the attribute sets may differ greatly from each other. The application of this measure is still under investigation.

Regarding the detailed results of the acoustics of the halls and the listening positions, an interested reader is also referred to the paper [68]. However, in short, the results suggest clear differences between the halls, the Konservatorio being the most reverberant and the Sello performing best in terms of definition. Also, the different listening positions are quite well separated in the perceptual space - for example, the close positions are evaluted to be perceptually louder as well as closer to the listener - while the different musical extracts give only slightly different sensory profiles. Distance and loudness attributes were apparently easiest to rate the samples with according to the number of these attributes and the high level of consensus. Almost every assessor also had one attribute related to reverberance and one attribute related to definition or clarity. Spatial aspects such as width, broadness and source width were also described by many and there were some timbral attributes, which were however spread out in several clusters. There were no attributes on the overall quality or preference, just as it was instructed.

In overall, this study proved that the individual vocabulary profiling method is feasible in studying the subjective perception of concert hall acoustics and yields results with great detail. Although, considering resources, this methodology is much lighter than a corresponding consensus vocabulary methods, it still requires quite a lot of time and commitment from both the experimenter and the test subjects. Also, the parallel comparison of the acoustical samples required the use of the loudspeaker orchestra, which is a special and advanced system developed during years of research. In this respect, it is difficult to reproduce this kind of experiment in the same scale without a corresponding system.

The aptitude of assessors to the descriptive analysis can be checked with a task specific screening procedure. The inexperienced assessors in this study showed clear differences in their behaviour in terms of the individual sensory profiles. However, at least five perceptual dimension were identified in the global analysis: 1. reverberance related to the size of space, 2. enveloping reverberance, 3. apparent source width, 4. loudness/distance, and 5. definition. These findings are also very well inline with the previous studies of concert hall acoustics.

Regarding prospective future studies, the work includes, among others, refining the test procedure to facilitate the task for the test subjects and further development of the loudspeaker orchestra in order to produce even better sounding stimuli. There is also the question about the correspondence and relation between the subjective perception and the objective measures, that should be attended more in detail. Also, investigating listeners' preferences and attentional focal points should be beneficial in gaining a better understanding of the overall perceptual experience of the acoustics inside a concert hall. If possible, it would obviously be very interesting to include the world's most appraised concert halls in such research. Well, the newly build Helsinki Music House is expected be a good start.

5 Summary

This thesis focused on the subjective perception of concert hall acoustics. The most important contribution of this work is the detailed presentation of the application of individual vocabulary profiling to the investigation of the perceptual characteristics of three Finnish concert halls. It was shown that this methodology can be well adapted to the study of acoustics and it yields versatile and valuable results. The behaviour of assessors was investigated with the individual sensory profiles obtained with Principal Component Analysis. This analysis was complemented with evaluating the complexity of the profiles with *beta*-coefficient. The results indicated clear differences between assessors in their ability to produce a set of attributes, which is descriptive and discriminative in many perceptual dimensions.

Additionally, this thesis included an extensive literature study of the state of concert hall acoustics research shedding light on the various unsolved aspects of relation between the objective acoustical measures and the subjective perception. The discussion about the research methodology in these studies highlighted a need for the development of more elaborate experimental designs. This discussion was complemented with the literature study on the descriptive analysis methods, what gave insight on the background and development of methods, which are popular in the field of sensory science, but might be more unknown to the researchers and engineers in the field of acoustics. It is important to be aware of these various approaches as well as their advantages as drawbacks in order to apply them in the audio field. This work offers a basic knowledge in this respect.

Finally, it is important to note that the employment of the virtual loudspeaker orchestra allowed the creation of the stimuli which could be simultaneously compared and evaluated. This evaluation technique lies at the core of the applied IVP method, which could not have been implemented without the current recording and signal processing techniques. That said, this study is a premium example of how the advancements of recording and processing techniques combined with novel sensory test methodology can produce new knowledge on our perception of acoustics and the listening experience in concert halls. Undoubtedly there is still a lot more to be discovered.

References

[1] N.W. Adelman-Larsen, E.R. Thompson, and A.C. Gade. Acoustics in rock and pop music halls. In *The 122nd Audio Engineering Society Convention*, Vienna, Austria, 2007. Paper 7141.

[2] Y. Ando. Subjective preference in relation to objective parameters of music sound fields with a single echo. *The Journal of the Acoustical Society of America*, 62:1436–1441, 1977.

[3] Y. Ando and M. Imamura. Subjective preference tests for sound fields in concert halls simulated by the aid of a computer. *Journal of Sound and Vibration*, 65(2):229–239, 1979.

[4] Y. Ando, H. Sakai, and S. Sato. Formulae describing subjective attributes for sound fields based on a model of the auditory-brain system. *Journal of Sound and Vibration*, 232(1):101–127, 2000.

[5] Y. Ando, S. Sato, T. Nakajima, and M. Sakurai. Acoustic design of a concert hall applying the theory of subjective preference, and the acoustic measurement after construction. *Acta Acustica united with Acustica*, 83(4):635–643, 1997.

[6] Yoichi Ando. A theory for individual preference of designing the sound field in a concert hall. *The Journal of the Acoustical Society of America*, 90(4):2238–2238, 1991.

[7] M. Barron. Subjective study of british symphony concert halls. *Acustica*, 66(1):1–14, 1988.

[8] M. Barron. *Auditorium acoustics and architectural design.* Verlag E & FS SPON, London, 1993.

[9] M. Barron and A.H. Marshall. Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration*, 77(2):211 – 232, 1981.

[10] S. Bayarri, I. Carbonell, L. Izquierdo, and A. Tárrega. Replicated triangle and duo–trio tests: Discrimination capacity of assessors evaluated by Bayesian rule. *Food Quality and Preference*, 19(5):519–523, 2008.

[11] S. Bech and N. Zacharov. *Perceptual Audio Evaluation. Theory, Method and Application.* John Wiley & Sons Ltd, 2006.

[12] T. Bech-Larsen and N. Asger Nielsen. A comparison of five elicitation techniques for elicitation of attributes of low involvement products. *Journal of Economic Psychology*, 20(3):315 – 341, 1999.

[13] L. Beranek. *Concert halls and opera houses.* Springer, 2004.

[14] L. Beranek. Concert hall acoustics–2008. *Journal of the Audio Engineering Society*, 56(7/8), 2008.

[15] L.L. Beranek. Concert hall acoustics-1992. *The Journal of the Acoustical Society of America*, 92:1, 1992.

[16] J. Berg. Opaque-a tool for the elicitation and grading of audio quality attributes. In *Proceedings of the 118th Convention of the Audio Engineering Society*, Barcelona, Spain, 2005. Paper 6480.

[17] J. Berg and F. Rumsey. Spatial attribute identification and scaling by repertory grid technique and other methods. In *Proceedings of the 16th International Audio Engineering Society Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, 1999. Paper 16–005.

[18] J. Berg and F. Rumsey. Identification of quality attributes of spatial audio by repertory grid technique. *Journal of the Audio Engineering Society*, 54(5):365, 2006.

[19] J. Bi. *Sensory discrimination tests and measurements: statistical principles, procedures, and tables*. Blackwell Pub, 2006.

[20] T.L. Bonebright. Perceptual structure of everyday sounds: A multidimensional scaling approach. In *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, 2001.

[21] J.S. Bradley. Contemporary approaches to evaluating auditorium acoustics. In *Proceedings of The 8th International Audio Engineering Society Conference on The Sound of Audio*, Washington D.C., USA, 1990. Paper 8–010.

[22] E. Breivik and M. Supphellen. Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques. *Journal of Economic Psychology*, 24(1):77 – 98, 2003.

[23] P. B. Brockhoff and P. Schlich. Handling replications in discrimination tests. *Food Quality and Preference*, 9(5):303–312, 1998.

[24] P.B. Brockhoff. Statistical testing of individual differences in sensory profiling. *Food quality and preference*, 14(5-6):425–434, 2003.

[25] T. Brookes and F. Kassier, R.and Rumsey. Training Versus Practice in Spatial Audio Attribute Evaluation Tasks. In *Proceedings of the 122nd Convention of the Audio Engineering Society*, Vienna, Austria, 2007. Paper 7117.

[26] S.E. Cairncross and L.B. Sjostrom. Flavor profiles-a new approach to flavor problems. *Food Technology*, 4(8):308–315, 1950.

[27] R. Cartier, A. Rytz, A. Lecomte, F. Poblete, J. Krystlik, E. Belin, and N. Martin. Sorting procedure as an alternative to quantitative descriptive analysis to obtain a product sensory map. *Food Quality and Preference*, 17(7-8):562–571, 2006.

[28] S. Choisel and F. Wickelmaier. Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. *Journal of the Audio Engineering Society*, 54(9):815–826, 2006.

[29] V. Dairou and J.M. Sieffermann. A comparison of 14 jams characterized by conventional profile and a quick original method, the flash profile. *Journal of Food Science*, 67(2):826–834, 2002.

[30] D. Davis. The role of Initial Time-Delay Gap in the Acoustic Design of Control Rooms for Recording and Reinforcing Systems. In *Proceedings of 64th International Convention of the Audio Engineering Society*, New York, USA, 1979. Paper 1547.

[31] J. Delarue and J.M. Sieffermann. Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food quality and preference*, 15(4):383–392, 2004.

[32] G. Dijksterhuis. Procrustes analysis in sensory research. *Data Handling in Science and Technology*, 16:185–219, 1996.

[33] K. Duineveld and M. Meyners. Hierarchical bayesian analysis of true discrimination rates in replicated triangle tests. *Food Quality and Preference*, 19(3):292–305, 2008.

[34] B. Escofier and J. Pagès. Multiple factor analysis (afmult package). *Computational statistics & data analysis*, 18(1):121–140, 1994.

[35] B.R. Gaines and M.L.G. Shaw. Knowledge acquisition tools based on personal construct psychology. *The knowledge engineering review*, 8(01):49–85, 1993.

[36] J.C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[37] D. Griesinger. Home page. http://www.davidgriesinger.com/. Accessed: 12/08/2011.

[38] D. Griesinger. Measures of spatial impression and reverberance based on the physiology of human hearing. In *Proceedings of the 11th International Audio Engineering Society Conference*, Portland, Oregon, USA, 1992. Paper 11–016.

[39] D. Griesinger. Room impression, reverberance, and warmth in rooms and halls. In *Proceedings of the 93rd International Convention of the Audio Engineering Society*, San Francisco, USA, 1992. Paper 3383.

[40] D. Griesinger. Spaciousness and envelopment in musical acoustics. In *Proceedings of the 101st Convention of the Audio Engineering Society*, Los Angeles, USA, 1996. Paper 4401.

[41] D. Griesinger. The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. *Acustica*, 83(4):721–731, 1997.

[42] D. Griesinger. General overview of spatial impression, envelopment, localization, and externalization. In *Proceedings of the 15th International Audio Engineering Society Conference*, Copenhagen, Denmark, 1998. Paper 15–013.

[43] D. Griesinger. Concert Hall Acoustics and Audience Perception [Applications Corner]. *IEEE Signal Processing Magazine*, 24(2):126–131, 2007.

[44] David Griesinger. Spatial impression and envelopment in small rooms. In *Proceedings of the 103rd Convention of the Audio Engineering Society*, New York, USA, 1997. Paper 4638.

[45] C. Guastavino and B.F.G. Katz. Perceptual evaluation of multi-dimensional spatial audio reproduction. *The Journal of the Acoustical Society of America*, 116:1105–1115, 2004.

[46] R. J. Hawkes and H. Douglas. Subjective acoustics experience in concert auditorio. *Acustica*, 24:235–250, 1971.

[47] J.R. Hurley and R.B. Cattell. The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7(2):258–262, 1962.

[48] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

[49] D. Isherwood, G. Lorho, V.-V. Mattila, and N. Zacharov. Augmentation, application and verification of the generalized listener selection procedure. In *the 115th Convention of the Audio Engineering Society*, New York, NY, USA, 2003. Paper 5894.

[50] ISO 8253-1:1989. *Acoustics – Audiometric test methods – Part 1: Basic pure-tone air and bone conduction threshold audiometry.* International Standards Organization, 1989.

[51] F.R. Jack and JR Piggott. Free choice profiling in consumer research. *Food quality and preference*, 3(3):129–134, 1992.

[52] I.T. Jolliffe. *Principal component analysis*. Springer Verlag, 2002.

[53] V.L. Jordan. Acoustical criteria for auditoriums and their relation to model techniques. *The Journal of the Acoustical Society of America*, 47:408–412, 1970.

[54] E. Kahle and M. Bruneau. *Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras / Validation of an Objective Model of the Perception of Room Acoustical Quality in an Ensemble of Concert Halls and Operas*. PhD thesis, 1995.

[55] G. Kelly. *The Psychology of Personal Constructs*. Norton, 1955.

[56] S. Kim and W.L. Martens. Verbal Elicitation and Scale Construction for Evaluating Perceptual Differences between Four Multichannel Microphone Techniques. In *Proceedings of the 122nd Convention of the Audio Engineering Society*, Vienna, Austria, 2007. Paper 7043.

[57] J. Kunert and M. Meyners. On the triangle test with replications. *Food Quality and Preference*, 10(6):477–482, 1999.

[58] A. Kuusinen, H. Vertanen, and T. Lokki. Assessor selection and behavior in individual vocabulary profiling of concert hall acoustics. In *Proceedings of the 38th International Audio Engineering Society Conference on Sound Quality Evaluation*, Piteå, Sweden, 2010. Paper 7–1.

[59] D. Labbe, A. Rytz, and A. Hugi. Training is a critical step to obtain reliable product profiles in a real food industry context. *Food Quality and Preference*, 15(4):341–348, 2004.

[60] R. Lacatis, A. Gimenez, A. Barba Sevillano, S. Cerda, J. Romero, and R. Cibrian. Historical and chronological evolution of the concert hall acoustics parameters. *The Journal of the Acoustical Society of America*, 123(5):3198–3204, 2008.

[61] C. Lavandier. *Validation perceptive d'un modele objectif de caracterisation de la qualite acoustique des salles = Perceptive validation of an objective model for the characterization of the room acoustic quality*. PhD thesis, 1989.

[62] H.T. Lawless and H. Heymann. *Sensory evaluation of food: principles and practices*. Aspen Publishers, 1999.

[63] S. Lê, J. Josse, and F. Husson. Factominer: An r package for multivariate analysis. *Journal of statistical software*, 25(1):1–18, 2008.

[64] S. Le Dien and J. Pagès. Hierarchical multiple factor analysis: application to the comparison of sensory profiles. *Food quality and preference*, 14(5-6):397–403, 2003.

[65] S.V. Legarth and N. Zacharov. Assessor selection process for multisensory applications. In *Proceedings of the 126th International Convention of the Audio Engineering Society*, Munich, Germany, 2009. Paper 7788.

[66] P. Lehman. *Über die Ermittlung raumakustischer Kriterien und deren Zusammenhang mit subjektiven Beurteilungen der Hörsamkeit (On the ascertainment of room acoustical criteria and correlation of the same with subjctive assessments of the acoustic overall impression).* PhD thesis, Technical University of Berlin, 1976.

[67] T. Lokki, R. Kajastila, and T. Takala. Virtual acoustic spaces with multiple reverberation enhancement systems. In *Proceedings of the 30th International Audio Engineering Conference*, Saariselkä, Finland, 2007. Paper 10.

[68] T. Lokki, J. Pätynen, A. Kuusinen, H. Vertanen, and S. Tervo. Concert hall acoustics assessment with individually elicited attributes. *The Journal of the Acoustical Society of America*, 130:835–849, 2011.

[69] T. Lokki, H. Vertanen, A. Kuusinen, J. Pätynen, and S. Tervo. Auditorium acoustics assessment with sensory evaluation methods. In *Proceedings of International Symposium on Room Acoustics*, Melbourne, Autralia, 2010.

[70] T. Lokki, H. Vertanen, and S. Siltanen. Intuitive Hand Gestures in Measurement of the Perceived Size of an Auditory Image of a Symphony Orchestra. In *Proceedings of the 38th International Audio Engineering Society Conference on Sound Quality Evaluation*, Piteå, Sweden, 2010. Paper 4–1.

[71] G. Lorho. Individual vocabulary profiling of spatial enhancement system for stereo headphone reproduction. In *Proceedings of the 119th International Convention of the Audio Engineering Society*, New York, NY, USA, 2005. Paper 6629.

[72] G. Lorho. Perceptual evaluation of mobile multimedia loudspeakers. In *Proceeding of the 122nd International Convention of the Audio Engineering Society*, Vienna, Austria, 2007. Paper 7050.

[73] G. Lorho. *Perceived quality evaluation: An application to sound reproduction over headphones.* PhD thesis, Aalto University School of Science and Technology, 2010.

[74] R.J. Marshall and S.P.J. Kirby. Sensory measurement of food texture by free-choice profiling. *Journal of Sensory Studies*, 3(1):63–80, 1988.

[75] J.A. McEwan, J.S. Colwill, and D.M.H. Thomson. The application of two free-choice profile methods to investigate the sensory characteristics of chocolate. *Journal of Sensory Studies*, 3(4):271–286, 1989.

[76] J. Merimaa and W. Hess. Training of listeners for evaluation of spatial attributes of sound. In *Proceedings of the 117th Convention of the Audio Engineering Society*, San Francisco, CA, USA, 2004. Paper 6237.

[77] J. Meyer. Acoustics and the performance of music. *Applied Mathematics and Mechanics*, 1, 2009.

[78] M. Meyners and K. Duineveld. Approximating the distribution of discrimination rates in replicated difference tests using Bayesian rule. *Food Quality and Preference*, 19(1):135–138, 2008.

[79] F. Rumsey N. Ford and T. Nind. Subjective evaluation of perceived spatial differences in car audio systems using a graphical assessment language. *Proceedings of the 112th International Convention of the Audio Engineering Society*, Munich, Germany, 2002. Paper 5547.

[80] T. Næs and R. Solheim. Detection and interpretation of variation within and between assessors in sensory profiling. *Journal of sensory studies*, 6(3):159–177, 1991.

[81] T. Neher, F. Rumsey, and T. Brookes. Training of listeners for the evaluation of spatial sound reproduction. In *Proceedings of the 112th Convention of The Audio Engineering Society*, Munich, Germany, 2002. Paper 5584.

[82] M. O'Mahony. Who told you the triangle test was simple? *Food Quality and Preference*, 6(4):227–238, 1995.

[83] C.E. Osgood. The nature and measurement of meaning. *Psychological Bulletin*, 49(3):197–237, 1952.

[84] J. Pagès. Multiple factor analysis: main features and application to sensory data. *Revista Colombiana de Estadistica*, 27(1):1–26, 2004.

[85] J. Pätynen, V. Pulkki, and T. Lokki. Anechoic recording system for symphony orchestra. *Acta Acustica united with Acustica*, 94(6):856–865, 2008.

[86] N. Pineau. *La performance en analyse sensorielle, une approche base de données*. PhD thesis, Universite de Bourgogne, 2006.

[87] R. Quesnel. Timbral Ear Trainer: Adaptive, Interactive Training of Listening Skills for Evaluation of Timbre. Copenhagen, Denmark, 1996. Paper 4241.

[88] R. Ratman, D. L. Jones, B. C. Wheeler, Jr. W. D. O'Brien, C. R. Lansing, and A. S. Feng. Blind estimation of reverberation time. *The Journal of the Acoustical Society of America*, 114(5):2877–2892, 2003.

[89] W.C. Sabine. *Collected papers on acoustics*. Harvard University Press, 1922.

[90] P. Schlich. Defining and validating assessor compromises about product distances and attribute correlations. *Data Handling in Science and Technology*, 16:259–306, 1996.

[91] P. Schlich, N. Pineau, D. Brajon, and E. M. Qannari. Multivariate control of panel performances. In *Proceedings of 7th Sensometrics Meeting*, Davis, USA, 2004.

[92] M.R. Schroeder. New Method of Measuring Reverberation Time. *The Journal of the Acoustical Society of America*, 37:409, 1965.

[93] M.R. Schroeder, D. Gottlob, and KF Siebrasse. Comparative study of European concert halls: Correlation of subjective preference with geometric and acoustic parameters. *The Journal of the Acoustical Society of America*, 56(4):1195–1201, 1974.

[94] J.M. Sieffermann. Le profil flash: un outil rapide et innovant d'évaluation sensorielle descriptive. *AGORAL 2000, XIIèmes rencontres "L'innovation: De l'idée au succès"*, pages 335–340, 2000.

[95] H. Stone, J. Sidel, and S. Oliver. Sensory evaluation by quantitative descriptive analysis. *Food Technology*, 28(11):24–33, 1974.

[96] A.S. Szczesniak. Classification of textural characteristics. *Journal of Food Science*, 28(4):385–389, 1963.

[97] H. Tachibana, Y. Yamasaki, M. Morimoto, Y. Hirasawa, Z. Maekawa, and C. Posselt. Acoustic survey of auditoriums in Europe and Japan. *The Journal Acoustical Society of Japan.(E)*, 10:73–85, 1989.

[98] S. Tarea, J.M. Sieffermann, and G. Cuvelier. Use of Flash Profile to Build a Product Set for More Advanced Sensory Study. Application to the Study of the Texture of Particles Suspensions. In *The 12th World Food Congress*, 2003.

[99] O. Tomic, A. Nilsen, M. Martens, and T. Næs. Visualization of sensory profiling data for performance monitoring. *LWT-Food Science and Technology*, 40(2):262–269, 2007.

[100] J. Vilkamo, T. Lokki, and V. Pulkki. Directional audio coding: Virtual microphone based synthesis and subjective evaluation. *Journal of the Audio Engineering Society*, 57(9):709–724, 2009.

[101] I. Wakeling, M. Raats, and H. MacFIE. A new significance test for consensus in generalized procrustes analysis. *Journal of sensory studies*, 7(2):91–96, 1992.

[102] A.A. Williams and G.M. Arnold. A comparison of the aromas of six coffees characterised by conventional profiling, free-choice profiling and similarity scaling methods. *Journal of the Science of Food and Agriculture*, 36(3):204–214, 1985.

[103] A.A. Williams and S.P. Langron. The use of free-choice profiling for the evaluation of commercial ports. *Journal of the Science of Food and Agriculture*, 35(5):558–568, 1984.

[104] R. Xiong, K. Blot, JF Meullenet, and JM Dessirier. Permutation tests for generalized procrustes analysis. *Food Quality and Preference*, 19(2):146–155, 2008.

[105] N. Zacharov and K. Koivuniemi. Audio descriptive analysis & mapping of spatial sound displays. In *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, 2001.

[106] N. Zacharov and G. Lorho. What are the requirements of a listening panel for evaluating spatial audio quality. In *Proceedings of The International Workshop on Spatial Audio and Sensory Evaluation Techniques*, 2006.

[107] N. Zacharov and V.-V. Mattila. GLS – a generalised listener selection procedure. In *Proceedings of the 110th International Convention of the Audio Engineering Society*, Amsterdam, the Netherlands, 2001. Paper 5405.

A PCA: Correlation circles

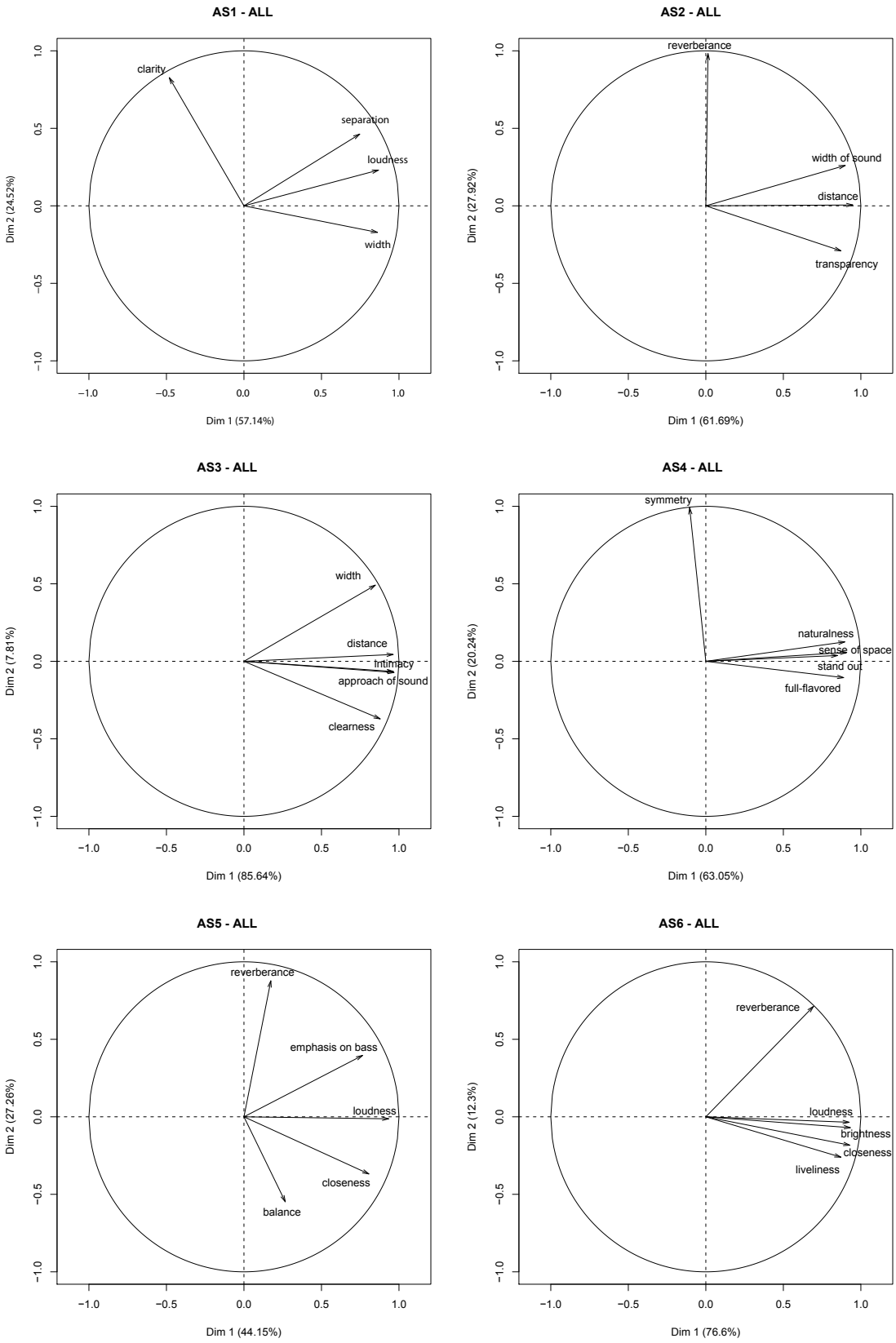


Figure A1: The attribute correlation circles in the latent spaces of the two first principal components. Assessors 1 - 6.

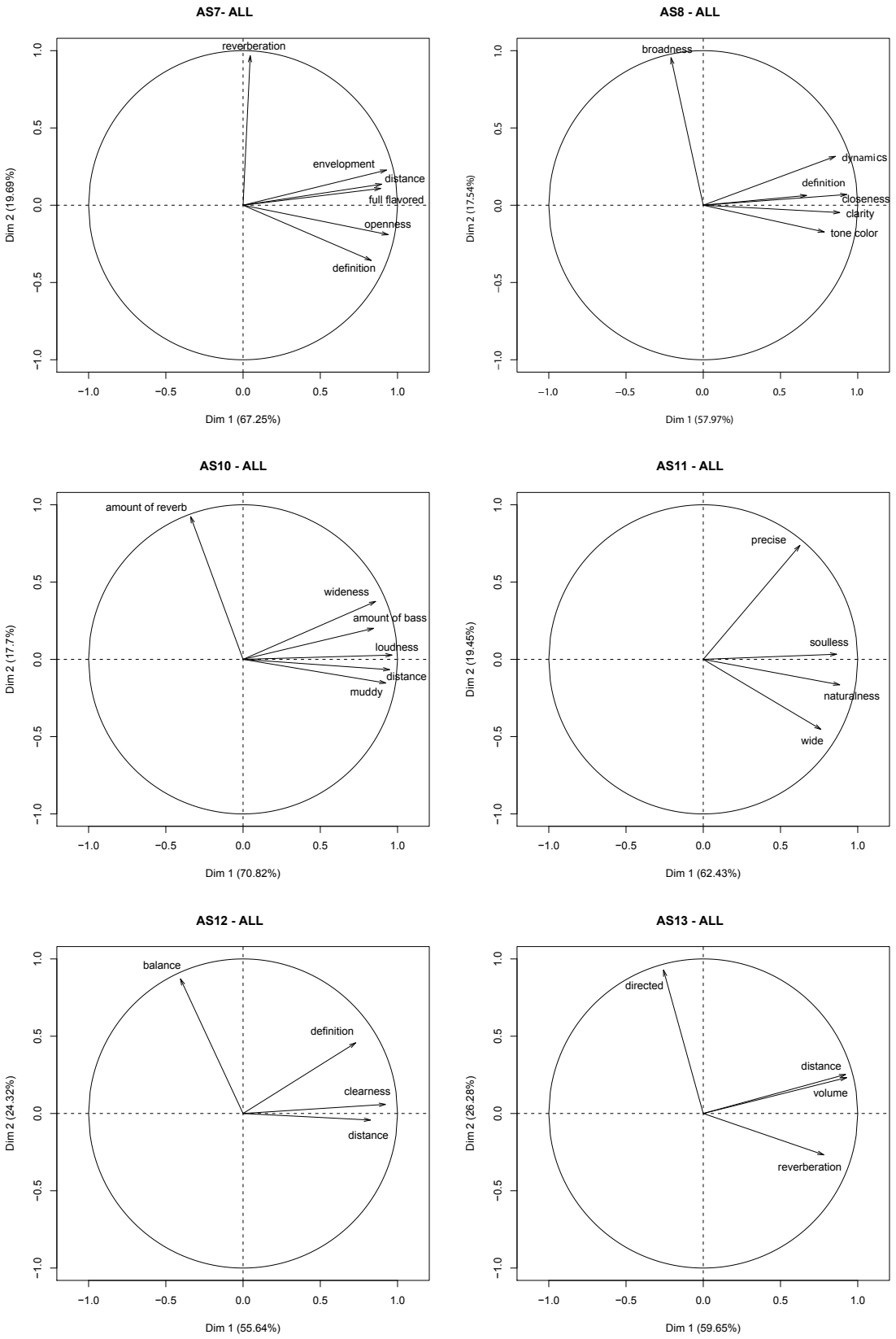


Figure A2: The attribute correlation circles in the latent spaces of the two first principal components. Assessors 7,8 and 10-13.

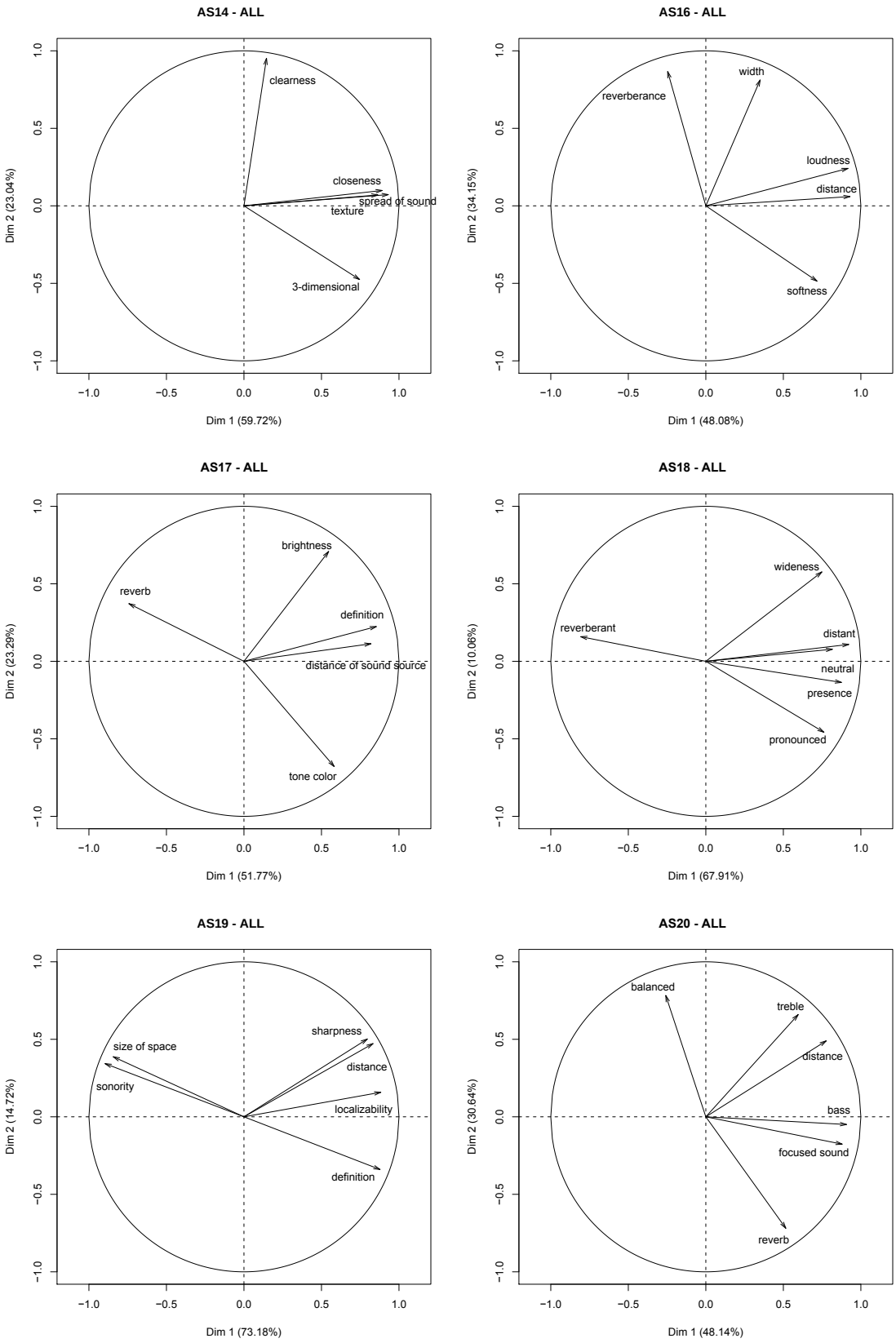


Figure A3: The attribute correlation circles in the latent spaces of the two first principal components. Assessors 14 and 16-20.

B Attributes

Table A1: All 102 attributes, with low and high anchors, elicited by 20 assessors (AS). Translated from Finnish. Finnish version of this table can be found in [58].

AS	Attribute	Low anchor	High anchor	AS	Attribute	Low anchor	High anchor
1	loudness (X37)	quiet	loud	11	precise (X12)	unclear	very precise
	brightness (X38)	dark	bright		wide (X13)	very narrow	very wide
	width of sound (X39)	narrow	wide		naturalness (X14)	unnatural	natural
	discrimination (X40)	blurry	clear		soulless (X15)	soulless	soulful
2	transparency (X22)	unbalanced	balanced	12	definition (X27)	difficult to define	easy to define
	width of sound (X25)	narrow	wide		distance (X28)	distant	near
	distance (X24)	distant	near		clearness (X30)	muddy	clear
	reverberance (X26)	dry	reverberant		balance (X29)	unbalanced	balanced
3	distance (X88)	far	near	13	volume (X47)	quiet	loud
	clearness (X89)	unclear	clear		distance (X48)	near	far
	intimacy (X90)	not intimate	intimate		reverberation (X50)	dry	wet
	approach of sound (X91)	reserved	aggressive		directed (X52)	directed	no clear direction
	width (X92)	narrow	wide				
4	naturalness (X7)	absorbed	natural	14	clearness (X16)	muffled	transparent
	full-flavored (X8)	thin	full		spread of sound (X17)	enveloping	piercing
	stand out (X9)	flat	distinct		closeness (X18)	far	close
	sense of space (X10)	narrow	spacious		texture (X19)	soft	hard
	symmetry (X11)	asymmetrical	symmetrical		3-dimensional (X20)	2D	3D
5	loudness (X2)	quiet	loud	15	distinctness (X59)	unclear	disctinct
	reverberance (X3)	unechoic	echoic		drr (X60)	without reverberation	reverberant
	closeness (X4)	distant	close		envelopment (X61)	kapea	laaja
	emphasis on bass (X5)	a little bass	a lot of bass		eq (X62)	thin	full
	balance (X6)	unbalanced	balanced		localizability (X63)	unclear	clear
6	liveliness (X64)	lifeless	lively	16	reverberance (X41)	little reverberant	very reverberant
	closeness (X65)	distant	close		softness (X42)	sharp	soft
	brightness (X66)	muddy	bright		loudness (X43)	quiet	loud
	reverberance (X67)	dry	reverberant		distance (X44)	far	near
	loudness (X69)	quiet	loud		width (X46)	narrow	wide
7	distance (X82)	far source	near	17	distance of sound	far	near
	envelopment (X83)	point source	enveloping		source (X32)		
	openness (X84)	stuffy	open		tone color (X33)	unequal	equal
	full-flavored (X85)	powerless	full		reverb (X34)	a little	a lot
	reverberation (X86)	dry	muddling		definition (X35)	messy	high definition
	definition (X87)	muddy	high definition		brightness (X36)	dark	bright
8	definition (X53)	blurry	definitive	18	distant (X76)	far	near
	closeness (X54)	far	near		reverberant(X77)	unclear	flat
	broadness (X55)	focused	broad		neutral (X78)	narrow	wide
	tone color (X56)	colored	balanced		pronounced (X79)	dark	bright
	dynamics (X57)	compressed	dynamic		wideness (X80)	condensed	wide
	clarity (X58)	muddy	clear		presence (X81)	far	present
9	depth (X70)	far	near	19	distance (X100)	far	near
	balance (X71)	cold	warm		localizability (X101)	can't locate	easy to locate
	intensity (X72)	muffled	loudness		definition (X102)	thickened	clear
	naturalness (X73)	unnatural	natural		sonority (X103)	dry	echoic
	breadth (X74)	mono	stereo		sharpness (X104)	cold	warm
	clearness (X75)	muddiness	clearness		size of the space (X105)	small	large
10	amount of reverb (X94)	dry	wet	20	reverb (X106)	little reverb	a lot of reverb
	wideness (X95)	condensed	broad		focused sound (X107)	point like	wide
	loudness(X96)	quiet	loud		distance (X108)	near	far
	distance (X97)	distant	close		bass (X109)	a lot of bass	little bass
	muddy (X98)	muddy	clear		treble (X110)	blocked	sharp
	amount of bass (X99)	low (a lot)	high (little)		balanced (X111)	muddy	distinct

C Matlab -code for *beta* -coefficient calculation

```
 %%

kh=dir('KH31_kaikki.txt');
for fno=1:length(kh)
    data1 = load(kh(fno).name);
end
%

%% GLOBAL
data = data1(2:37, 2:size(data1, 2));
attributes = size(data1, 2) - 1;

% CENTERING...
ave = sum(sum(data))/(size(data, 1) * size(data, 2)); % matrix average
for idx = 1:size(data, 1)
    for idx2 =  1:size(data, 2)
        data(idx, idx2) = data(idx, idx2) - ave;
    end;
end;

% Association matrix Wi and beta-coefficient
Wi = data * data';
beta_all = ((trace(Wi))^2)/(trace(Wi ^ 2));
Wii_all = trace(Wi ^ 2);

%% MOZART

data = data1(2:10, 2:size(data1, 2));

% CENTERING...

ave = sum(sum(data))/(size(data, 1) * size(data, 2)); % matrix average
for idx = 1:size(data, 1)
    for idx2 =  1:size(data, 2)
        data(idx, idx2) = data(idx, idx2) - ave;
    end;
end;

% Association matrix Wi and beta-coefficient

Wi = data * data';
beta_moz = ((trace(Wi))^2)/(trace(Wi ^ 2));
Wii_moz = trace(Wi ^ 2);

%% BEETHOVEN
data = data1(11:19, 2:size(data1, 2));
ave = sum(sum(data))/(size(data, 1) * size(data, 2)); % matrix average
for idx = 1:size(data, 1)
    for idx2 =  1:size(data, 2)
        data(idx, idx2) = data(idx, idx2) - ave;
```

```
51      end;
52  end;
53
54  % Association matrix Wi and beta-coefficient
55  Wi = data * data';
56  beta_bee = ((trace(Wi))^2)/(trace(Wi ^ 2));
57  Wii_bee = trace(Wi ^ 2);
58
59  %% BRUCKNER
60  data = data1(20:28, 2:size(data1, 2));
61  ave = sum(sum(data))/(size(data, 1) * size(data, 2)); % matrix average
62  for idx = 1:size(data, 1)
63      for idx2 =  1:size(data, 2)
64          data(idx, idx2) = data(idx, idx2) - ave;
65      end;
66  end;
67
68  % Association matrix Wi and beta-coefficient
69  Wi = data * data';
70  beta_bru = ((trace(Wi))^2)/(trace(Wi ^ 2));
71  Wii_bru = trace(Wi ^ 2);
72
73  %% MAHLER
74  data = data1(29:37, 2:size(data1, 2));
75  ave = sum(sum(data))/(size(data, 1) * size(data, 2)); % matrix average
76  for idx = 1:size(data, 1)
77      for idx2 =  1:size(data, 2)
78          data(idx, idx2) = data(idx, idx2) - ave;
79      end;
80  end;
81
82  % Association matrix Wi and beta-coefficient
83  Wi = data * data';
84  beta_mah = ((trace(Wi))^2)/(trace(Wi ^ 2));
85  Wii_mah = trace(Wi ^ 2);
86
87  %%
```