Aalto University
School of Science and Technology
Department of Signal Processing and Acoustics
Espoo 2010 Report 20

PERCEIVED QUALITY EVALUATION AN APPLICATION TO SOUND REPRODUCTION OVER HEADPHONES

Gaëtan Lorho

Aalto University
School of Science
and Technology

Aalto University
School of Science and Technology
Department of Signal Processing and Acoustics
Espoo 2010 Report 20

PERCEIVED QUALITY EVALUATION
AN APPLICATION TO SOUND REPRODUCTION OVER HEADPHONES

Gaëtan Lorho

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Electronics, Communications and Automation for public examination and debate in Auditorium S4 at the Aalto University School of Science and Technology (Espoo, Finland) on the 8th of June 2010, at 12 noon.

Aalto University
School of Science and Technology
Faculty of Electronics, Communications and Automation
Department of Signal Processing and Acoustics

Distribution:
Aalto University
School of Science and Technology
Faculty of Electronics, Communications and Automation
Department of Signal Processing and Acoustics
P.O. Box 13000
FI-00076 AALTO
Tel. +358 9 47001
Fax +358 9 452 3614
E-mail heidi.koponen@tkk.fi

ISBN 978-952-60-3195-8 (Printed)
ISBN 978-952-60-3196-5 (Electronic)
ISSN 1797-4267

Multiprint Oy
Espoo, Finland 2010

Aalto University

ABSTRACT OF DOCTORAL DISSERTATION	AALTO UNIVERSITY SCHOOL OF SCIENCE AND TECHNOLOGY P.O. BOX 11000, FI-00076 AALTO http://www.aalto.fi

Author Gaëtan Lorho

Name of the dissertation
Perceived quality evaluation: An application to sound reproduction over headphones

Manuscript submitted 15.2.2010	Manuscript revised 24.5.2010
Date of the defence 8.6.2010	
☒ Monograph	☐ Article dissertation (summary + original articles)

Faculty	Faculty of Electronics, Communications and Automation
Department	Department of Signal Processing and Acoustics
Field of research	Acoustics
Opponent(s)	Dr. Brian Katz and Prof. El Mostafa Qannari
Supervisor	Prof. Matti Karjalainen
Instructor	Dr. Nick Zacharov

Abstract

Quality evaluation of sound reproduction systems is ultimately a perceptual matter and is intuitively associated to the hedonic domain with questions regarding the goodness of a given system. However, beyond this integrative evaluation approach relying on a measure of overall impression, quality can also be tackled from a more elemental and analytical viewpoint with the aim of understanding the perceptual components affecting quality. The present thesis work focuses on this topic referred to as 'sensory analysis' and explores more specifically two different types of sensory techniques using quantitative scales associated to verbal descriptors to characterize stimuli: the consensus vocabulary and individual vocabulary methods.

Sound reproduction over headphones is the application considered in this investigation on sensory analysis. Advanced applications for headphone rendering of spatial or three-dimensional sound and virtual surround sound have been developed but the perceived quality evaluation of such systems remains an issue. In the present work, the specific case of spatial enhancement algorithms for the reproduction of music or movie sound over headphones is addressed. Firstly, the development of a consensus vocabulary of headphone sound perception is reported in which a panel of screened assessors created a consensual set of sixteen attributes with their associated definition, word anchors and audio exemplars to describe the perceptual characteristics of a wide range of headphone sound stimuli. The applicability of this vocabulary is tested on two sets of headphone sound samples. Secondly, a procedure entitled 'individual vocabulary profiling' developed for rapid sensory evaluation of audio applications by inexperienced assessors using an individual elicitation method is presented and its application is illustrated on a set of spatial enhancement systems. Lastly, these two approaches are compared to highlight their respective benefits, challenges and limitations and their suitability for different sensory evaluation projects.

Throughout this thesis, statistical techniques for the analysis of sensory data are also investigated. The attribute rating data resulting from the two sensory profiling experiments is analyzed systematically. Several advanced multivariate data analysis methods are explored and applied to these data sets including generalized Procrustes analysis, multiple factor analysis, hierarchical multiple factor analysis and parallel factor analysis.

The concepts and techniques presented in this thesis are applicable to many audio quality problems and can be adapted to other perceptual domains.

Keywords audio quality, spatial sound, sensory analysis, vocabulary development, multivariate data analysis

ISBN (printed)	978-952-60-3195-8	ISSN (printed)	1797-4267
ISBN (pdf)	978-952-60-3196-5	ISSN (pdf)	1797-4267
Language	English	Number of pages	248
Publisher	Aalto University School of Science and Technology		

☒ The dissertation can be read at http://lib.tkk.fi/Diss/2010/isbn9789526031965

Aalto-yliopisto

VÄITÖSKIRJAN TIIVISTELMÄ	AALTO-YLIOPISTO TEKNILLINEN KORKEAKOULU PL 11000, 00076 AALTO http://www.aalto.fi
Tekijä Gaëtan Lorho	
Väitöskirjan nimi Äänenlaadun arviointi kuulokekuuntelussa	
Käsikirjoituksen päivämäärä 15.2.2010	Korjatun käsikirjoituksen päivämäärä 24.5.2010
Väitöstilaisuuden ajankohta 8.6.2010	
☒ Monografia	☐ Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)

Tiedekunta	Elektroniikan, tietoliikenteen ja automaation tiedekunta
Laitos	Signaalinkäsittelyn ja akustiikan laitos
Tutkimusala	Akustiikka
Vastaväittäjä(t)	Dr. Brian Katz and Prof. El Mostafa Qannari
Työn valvoja	Prof. Matti Karjalainen
Työn ohjaaja	Dr. Nick Zacharov

Tiivistelmä

Äänentoistojärjestelmien laadun arvioinnissa on perimmiltään kyse aistinvaraisista havainnoista, jotka liittyvät intuitiivisesti käyttäjän mieltymyksiin ja kysymykseen äänentoiston hyvyydestä. Tämän kokonaisvaikutelman arviointiin käytetyn yhdenmukaisen lähestymistavan lisäksi laatuun voi ottaa perusteellisemman ja analyyttisemman näkökulman tavoitteena ymmärtää äänenlaatuun vaikuttavia aistinvaraisia tekijöitä. Tätä prosessia kutsutaan tässä väitöskirjassa ‘aistinvaraiseksi arvioinniksi’. Tutkimuksessa käytetään kahta erilaista aistinvaraista kuvailevaa menetelmää, joissa tutkittavia näytteitä kuvataan hyödyntäen kvantitatiivisia asteikkoja yhdistettynä sanallisiin kuvauksiin. Menetelmät ovat: kvantitatiivinen kuvaileva analyysi yhdistettynä yhteisesti sovittuihin ominaisuuksiin ja kvantitatiivinen kuvaileva analyysi yhdistettynä yksilöllisiin ominaisuuksiin.

Tutkimuskohteena tässä aistinvaraisessa arvioinnissa on kuulokkeiden äänentoisto-ominaisuudet. Kuulokeäänentoistoa varten on kehitetty tila- ja 3D-äänen prosessointimenetelmiä, mutta yleisesti näiden menetelmien äänenlaadun arviointi on jäänyt vaille huomiota. Tässä työssä keskitytään erityisesti tilaefektejä (stereo enhancement) tuottaviin algoritmeihin musiikin tai elokuvien äänentoistossa kuulokkeiden avulla. Ensimmäiseksi raportoidaan menetelmä, jossa arviointiin käytettiin yhteisesti sovittuja kuulokkeiden ääntä kuvaavia ominaisuuksia. Etukäteen testattu ja valittu raati loi 16 sanallisesti määriteltyä ominaisuutta, joista raadin jäsenet olivat yksimielisiä. Ominaisuuksille määriteltiin käytettyyn asteikkoon liittyen sanalliset ankkurit ja ääniesimerkit, jotka kuvasivat kuulokkeiden äänentoiston aistinvaraisia ominaisuuksia monipuolisesti. Tämän yhteisesti kehitetyn ja sovitun sanaston käyttökelpoisuutta testattiin kahdella ääninäytejoukolla. Toiseksi raportoidaan menetelmä, jota kutsutaan ”yksilölliseksi kuvailevaksi menetelmäksi”. Menetelmää sovellettiin tilaääniefektien tutkimiseen ja se on kehitetty nopeaan aistinvaraiseen arviointiin, jossa arvioijia ei tarvitse kouluttaa, ja arvioijat käyttävät itse luomiaan sanastoja.

Tutkimuksissa käytetään tilastollisia menetelmiä aistinvaraisten aineistojen analysoinnissa. Edellä mainittuja kahta menetelmää verrataan, korostamalla menetelmien etuja, haasteita ja rajoitteita. Vertailun tarkoituksena on selvittää menetelmien soveltuvuus erilaisiin aistinvaraisiin arviointiprojekteihin. Ominaisuuksien voimakkuusarviot (kvantitatiivinen asteikko) molemmista aistinvaraisista tutkimuksista on analysoitu systemaattisesti. Analyysissä sovellettiin useita kehittyneitä monimuuttujamenetelmiä, Prokrustes-analyysiä, monimuuttujafaktorianalyysia, hierarkkista monimuuttujafaktorianalyysiä, ja yhdensuuntaista faktorianalyysia.

Väitöskirjassa esitetyt konseptit ja tekniikat soveltuvat moniin äänentoiston laadunarviointiongelmiin ja niitä voidaan soveltaa myös muilla aloilla.

Asiasanat	tilaääni, aistinvarainen arviointi, sanaston kehittäminen, tilastolliset monimuuttujamenetelmät		
ISBN (painettu)	978-952-60-3195-8	ISSN (painettu)	1797-4267
ISBN (pdf)	978-952-60-3196-5	ISSN (pdf)	1797-4267
Kieli	Englanti	Sivumäärä	248
Julkaisija	Aalto-yliopisto Teknillinen korkeakoulu		
☒ Luettavissa verkossa osoitteessa http://lib.tkk.fi/Diss/2010/isbn9789526031965			

Preface

The research reported in this thesis has been carried out at Nokia during the period of 2002-2009. The initiative and the experimental part of this work took place over the first four years but writing the monograph turned out to be a slightly longer 'side project' than expected. I believe this thesis would not have been completed without the continuous encouragements of my supervisor Prof. Matti Karjalainen and my tutor Dr. Nick Zacharov. I feel very privileged to have carried out this work with this ideal supervising team. The deep knowledge, enthusiasm, spontaneity, positive feedback and patience of Matti and Nick were the essential drivers that strengthened my determination to complete this thesis project.

This dissertation is the result of a kind of initiatory journey and along the way I had the chance to meet and interact with many people from different research horizons. This project can be summarized in four learning phases involving the research fields of perceptual audio and sensory science.

The first phase of this journey relates to perceptual audio and took place during the period of 2002-2003 at Nokia Research Center, Speech and Audio Systems Laboratory. First, I would like to thank Dr. Nick Zacharov and Dr. Ville-Veikko Mattila for their work which served as inspiration for the topic of this thesis. Without their pioneering research in the field of perceptual evaluation of audio at Nokia, the present work would not have been possible. I am also thankful to Dr. Jyri Huopaniemi for supporting my efforts before and during this period. His visionary drive and enthusiasm had a real impact on my will to get started with this thesis project. In addition, I would like to thank all my colleagues in Helsinki and in Tampere who created a very inspiring environment for audio research at NRC. This includes Mr. Jarmo Hiipakka, Mr. Jussi Virolainen, Mr. Timo Sorsa, Mr. Matti Paavola, Mr. Tommi Keranen, Mr. Olli Tuomi, Mr. Juha Salmela, Mr. Miikka Vilermo and many others. Dr. Ole Kirkeby deserves special thanks for challenging me with listening test ideas to evaluate spatial audio enhancement systems. I would also like to acknowledge Mr. Kalle Koivuniemi for his contribution to the consensus vocabulary development work described in this thesis, Mr. David Isherwood for our fruitful collaboration on the topic of perceptual evaluation of audio, and Mr. Henri Toukomaa for running the experiments reported in Chapter 6 of this thesis.

The second phase can be described as a transition in the project and occurred through a series of sensory evaluation workshops held in 2003. This period of interaction with experts from the fields of audio science and food science has been really insightful for me and I would like to thank all the members of this workshop team: Dr.

Jannie Vestergaard, Dr. Ulla Suhonen, Ms. Soili Lampolahti, Dr. Michael Bom Frøst, Dr. Søren Bech, Dr. Geoffrey Martin, Mr. Søren Jørgensen, Dr. Nick Zacharov, Dr. Ville-Veikko Mattila, Mr. Kalle Koivuniemi and Mr. David Isherwood.

In the third phase of this project, I became familiar with sensory science during a 6-month visit at the Department of Food Science, Copenhagen University in 2003/2004. This academic stay in Denmark shaped the content of my thesis work. I am very grateful to Prof. Magni Martens and Prof. Wender Bredie for welcoming me in the Sensory Science Group. During this time, I was able to complete the experimental work of my thesis and I learned a lot about food science. I really had a great time at the University of Copenhagen and would like to thank all the members of the Sensory Science group for sharing their time, expertise and ideas. In addition, I had the great privilege to meet Prof. Rasmus Bro when following the Advanced Chemometrics course held by the Chemometrics Group in the Department of Food Science. Rasmus deserves my sincere gratitude for introducing me to the world of multi-way data analysis. His knowledge and enthusiasm have been a great source of motivation for me to learn more about this research topic.

Back from Denmark, I was able to apply some of my acquired knowledge to the field of audio and I want to acknowledge Mr. David Pegon-Johnson for his interest in applying to a different sensory modality the evaluation methods presented in this work. In 2006, the last phase of this journey started, that is, the slow and solitary process of writing the thesis monograph. I would like to thank Ms. Suzanne Luoto from the Helsinki information team at Nokia for digging out rare and old references and providing me with many interesting publications to read during this writing phase. Many thanks to Mr. Olli Niemi for his constant support and patience regarding the completion of my thesis monograph. Thanks also to all the members of the authentic 'AQUA' team at Nokia Devices R&D, including my former colleagues in Helsinki: Mr. Markus Vaalgamaa, Mr. Mika Hanski and Mr. Marko Takanen.

I am grateful to the pre-examiners of this thesis, Prof. El Mostafa Qannari and Dr. Sylvain Choisel, for their helpful comments and positive feedback on the manuscript. Thanks also to Dr. Ulla Suhonen, Dr. Tapio Loki and Mr. Jussi Virolainen for taking care of the Finnish translation of the thesis abstract and to Ms. Heidi Koponen for helping me with the preparation of the defense.

I also wish to thank my friends for being around during this long process and for their ability or attempts to keep my mind away from this thesis work: Jean-Luc with sports, Martin with culture, Arnaud with cinema, Antoine with socializing and all the others. My special thanks go to Sanna who supported me during this work. Finally, I would like to dedicate this work to my family in Brittany (France).

The 23rd of May 2010 in Kulosaari (Helsinki, Finland)

Gaëtan Lorho

Contents

List of Abbreviations

ADAM	Audio Descriptive Analysis and Mapping
AHC	Agglomerative hierarchical clustering
ANOVA	Analysis of variance
BRIR	Binaural room impulse response
CI	Confidence interval
CV	Consensus vocabulary
CVA	Canonical variate analysis
DF	Diffuse-field
FCP	Free-Choice Profiling
FF	Free-field
FP	Flash Profile
GLS	Generalized Listener Selection
GPA	Generalized Procrustes analysis
HMFA	Hierarchical multiple factor analysis
HRTF	Head-related transfer function
IV	Individual vocabulary
IVP	Individual Vocabulary Profiling
MANOVA	Multivariate analysis of variance
MDS	Multidimensional scaling
MFA	Multiple factor analysis
PARAFAC	Parallel factor analysis
PCA, PC	Principal component analysis, Principal component
PSA	Perceptual Structure Analysis
QDA	Quantitative Descriptive Analysis
QFP	Quantitative Flavor Profiling
RGT	Repertory Grid Technique
WFS	Wave field synthesis

Chapter 1

Introduction

1.1 Background

The first part of this thesis title, 'perceived quality evaluation' combines three important elements of engineering. Evaluation is an essential part of any system or process control and optimization but evaluating complex systems such as those designed for human beings can be a difficult task. The term 'Quality' refers intuitively to performance but this word relates to a rather broad and abstract concept and it is usually associated to hedonic elements when it concerns systems involving sensory modalities, e.g. audio or/and visual systems. Perception is a central component in quality when dealing with systems interacting with any sensory modality or the combination of those. Perceived quality evaluation can be simply summed up to the measure of performance of a (usually complex) system by subjects interacting with it.

The development of sound reproduction techniques brought multichannel sound to cinema theaters in the 50's and has been recently adapted for home applications with the introduction of 5.1 loudspeaker systems and the availability of movie sound, music and gaming sound material in multichannel audio format. While loudspeaker reproduction systems can include any number of loudspeakers, headphones are usually bound to two channels, i.e. one channel per ear, which is problematic for the replay of multichannel sound material. In addition, the acoustic experience resulting from such a sound reproduction scenario is very unnatural or artificial and has specific spatial imaging characteristics. Basic research on the psychoacoustics of headphones has been reported in the literature and advanced applications for headphone rendering of spatial or three-dimensional sound and virtual surround sound have been developed. However, the perceived quality evaluation and optimization of such systems remains an issue.

Headphone sound reproduction is the application context in which the research topic of perceived quality evaluation was investigated in the present doctoral thesis. In practical terms, the author looked for suitable test methodologies for the reliable measurement of the perceived characteristics of audio systems. This type of measurement implies naturally human subjects but an essential component of this work was to turn a panel of subjects into an objective measurement system.

1.2 Scope of the thesis

The research reported in this thesis relates broadly to the topic of sound quality evaluation and comprises a set of perceptual studies on spatial sound reproduction over headphones. Concepts behind perceived quality are investigated before concentrating on the more specific topic of sensory analysis, that is, the elicitation of objective responses to the properties of a stimulus as perceived by the human senses. A large number of sensory analysis techniques can be found in the literature and one aim of this thesis was to classify such methods and assess their suitability for the perceptual evaluation of audio. The main focus of the present work is on verbal descriptive analysis which is a class of sensory methods using quantitative scales associated to verbal descriptors to characterize stimuli. This type of perceptual evaluation is exploratory in nature and represents one of the most sophisticated tools in sensory science. It has been developed and applied extensively in the field of food science and has also some applications in other fields of research including audio quality. Two different routes exist for verbal descriptive analysis, that is, the consensus vocabulary approach using a panel of assessors to develop a common set of sensory descriptors and the individual vocabulary approach letting each assessor of the panel develop his or her own set of descriptors. Based on previous work in the field of sensory science and perceptual audio evaluation, the author explored systematically these two approaches and adapted them with the aim of obtaining rapid methods for the perceptual evaluation of audio applications.

The exploration of the consensus and individual approaches to verbal descriptive analysis considered in this doctoral thesis includes a practical application to headphone sound reproduction. Two listening experiments have been carried out to evaluate a large set of stimuli relating to spatial enhancement systems for music reproduction over headphones. An additional research topic investigated in the present thesis relates to the data resulting from this type of experiments. Multivariate data analysis is an important tool in sensory analysis and the author explored several advanced analysis techniques and applied them to the experimental data resulting from the two audio experiments of this thesis.

1.3 Contents of the thesis

This doctoral thesis comprises nine principal chapters and is structured in four main parts as follows.

The first part (Chapter 2) gives a broad overview of perceived quality evaluation. Firstly, a definition of quality is proposed and a structured framework for perceived quality evaluation is developed in which sensory analysis has a central role. Secondly, aspects of sensory analysis are developed with an emphasis on verbal elicitation methods. The second part (Chapter 3) provides a short overview of sound recording/reproduction methods for loudspeaker and headphone applications and describes the perceptual aspects associated with these different scenarios. A literature review of perceptual evaluation studies is presented to give a broad perspective on

the research field of sound perception. The last part of this chapter illustrates the 'traditional' approach to sound quality evaluation by summarizing three preference studies carried out by the author on spatial enhancement algorithms for headphone sound reproduction.

Following these two background chapters, the core of this thesis work is organized in two larger parts reporting an investigation on two forms of verbal descriptive sensory analysis. The former part covers the consensus vocabulary approach in three chapters with firstly an extensive review of consensus vocabulary methods (Chapter 4), secondly an application of this approach to the development of a consensus vocabulary of spatial sound reproduction over headphones (Chapter 5) and thirdly an application of this vocabulary to the evaluation of spatial enhancement systems (Chapter 6). The latter part covers the individual vocabulary approach with a relatively similar structure including firstly an extensive review of individual vocabulary methods (Chapter 7), secondly the development of a procedure for individual vocabulary development (Chapter 8) and thirdly an application of this procedure to the evaluation of spatial enhancement systems (Chapter 9). These two large parts of the thesis are discussed in Chapter 10 with a comparative analysis of the results from the two experiments and a summary of the methodological differences between the two verbal descriptive analysis approaches.

1.4 Contributions of the author

The author's contribution in relation to this thesis combines elements of sound quality evaluation and sensory science and can be summarized as follows.

A novel structured framework for (sound) quality evaluation was proposed that combines concepts relating to perceived quality and sensory science (Chapter 3). The author investigated thoroughly the field of verbal descriptive sensory analysis and provided a comprehensive review of the concepts, techniques and implementation aspects of the two experimental methodologies available in this category, that is, the consensus vocabulary approach (Chapter 4) and the individual vocabulary approach (Chapter 7). This investigation included also a detailed presentation of concepts relating to sensory panel performance assessment for the two test methods. These two verbal descriptive sensory approaches have been applied to the field of audio in a systematic manner. The author established and applied a generic consensus vocabulary for spatial sound reproduced over headphones (Chapter 5 and 6) and developed of a novel rapid test procedure for sensory profiling using an individual elicitation approach (Chapter 8 and 9).

An important part of this research work relates to the analysis of sensory data. The author investigated systematically the attribute rating data resulting from the two sensory profiling experiments reported in this thesis. Especially, the individual vocabulary data (Chapter 9) was treated thoroughly to illustrate the specific of this data. Several advanced multivariate data analysis methods from the field of sensory science were explored and applied to the experimental data resulting from the two perceptual audio quality experiments. The use of multiple factor analysis (MFA) and hierarchical multiple factor analysis (HMFA) for the analysis of several sets of

multivariate data (Chapters 6 and 9 respectively) was demonstrated. The application of multi-way data analysis methods to sensory data was investigated and a novel application of PARAFAC2 that can handle the 4-way data structure of the individual vocabulary data generated in this thesis work was proposed. Through these different analyses, the author was able to identify and quantify the main perceptual differences between spatial enhancement systems for headphone reproduction.

Chapter 2

Overview of Perceived Quality Evaluation

2.1 Introduction

Quality can be defined in very different ways depending on the application of interest. An important aspect of this concept however is that it usually relates to the evaluation of an entity. Besides, the question of *perceived* quality highlights the relationship between this entity and a (human) subject. Perceptual evaluation is often limited to a single modality, e.g., audition or vision, but it can also involve a combination of modalities. In this chapter, concepts relating to quality are discussed with a focus on the auditory modality, an example of which could be the perceived quality of speech reproduced over the integrated loudspeaker of a mobile phone. It should be noted however that the ideas discussed below for the auditory modality would apply equally to another perceptual domain and can be extended to multimodal perceptual quality problems. An overview of the approaches to measure or characterize sound stimuli is presented in Section 2.2 with the aim of providing a frame of reference to the reader and positioning the work of the author within a larger context. This section also describes an approach to perceived quality evaluation referred to as 'preference mapping' and highlights the importance of sensory analysis in this quality evaluation framework. Section 2.3 focuses on sensory analysis, with first a discussion on matters relating to test subject classification and sensory panels, then an overview of existing sensory analysis methods and finally a more detailed presentation of descriptive analysis techniques employing the verbal elicitation approach together with some justifications of their use in the context of the research presented in this thesis.

2.2 Approaches to (sound) quality evaluation

2.2.1 Definitions of quality

Perceived audio quality evaluation is an area of audio engineering that has a broad range of applications from fundamental research on spatial sound perception (e.g. Rumsey, 2002) to more applied studies such as the evaluation of audio compression

algorithms (e.g. Soulodre *et al.*, 1998). The versatile use of the term *quality* in these different contexts indicates the loose meaning of what Blauert and Jekosch (2003) described as "*a mental construct which is often insufficiently defined and, consequently, not understood properly by many*". As the topic of the present work relates primarily to quality, some clarifications are needed regarding the definition of this term. Martens and Martens (2001) discussed the different meanings of quality and presented four common definitions for this word. This terminology is relevant in the context of sound evaluation and is therefore shortly reviewed below.

The first definition of quality presented by Martens and Martens (2001) relates to the inherent characteristics of an entity[1]. The focus in this case is placed on describing the properties of the entity in an objective sense. The second definition considers quality as a degree of excellence. This approach is more subjective and is somehow in conflict with the first definition. The meaning of excellence/goodness can also be problematic. The dangers of defining a degree of excellence from a single individual perspective, that is, based on the judgment of one person, are highlighted by the authors who oppose this classical approach to the market study approach which employs a large number of representative people to identify patterns of responses common to the whole group of subjects or specific to sub-groups. The third definition of quality is stated in ISO 9000 (2000) as "*the ability of a set of inherent characteristics of a product, system or process to fulfil requirements of costumers and other interested parties*". Quality refers in this case to a relational concept between the description of an entity and a set of stated or implied requirements (e.g. goodness) for this entity. To a certain extent, the objectivity and subjectivity aspects are linked through this third ISO definition. Finally, the fourth definition of quality considered by Martens and Martens (2001) is more abstract and relates to an individually experienced event or action.

Similar distinctions between quality definitions have been described in the context of perceived sound quality. For example, the difference made by Martens and Martens (2001) between the first definition dealing with a description of the characteristics of an entity and the second definition dealing with an evaluation of the goodness of the entity has also been highlighted by Letowski (1989), Martens and Zacharov (2003) and Blauert and Jekosch (2003). It can be noted that these publications employed the term *character* instead of *quality* to address the aspects relating specifically to the first definition. The ISO definition described above is the most relevant in the context of this thesis work and it shares similarities with the conceptual approach to sound quality proposed by Blauert and Jekosch (2003) which can be summarized in two steps as follows. The first step of the quality evaluation requires the determination of the character of the sound sample on the one hand and the character of the reference on the other hand. In the second step, a comparison between these two objective and quantitative character profiles is performed and a quality rating is assigned based on some measure of similarity. The specification of the reference is usually application

[1]An entity can be defined as the material or immaterial object under study. In the context of a perceptual study, this might refer to a 'product', a 'system' or a 'stimulus'.

specific and is considered by Blauert and Jekosch (2003) as a critical element of this process. The ISO definition of quality and the approach proposed by Blauert and Jekosch (2003) are related in the sense that they both rely on a comparison of characteristics between an entity and some target to be defined, i.e., the reference in the former case and the stated or implied requirements in the latter case.

The author adopts this idea of relational concept in the present work and proposes to define **Quality** as **a measure of the distance between the character of an entity under study and the character of a target associated with this entity**. Two comments have to be made at this point about this definition. Firstly, it should be noted that the specification of the target is considered as an integral part of the quality evaluation problem. In practice, this target is either known prior to the study or needs to be defined in the process of the study. In the example of an audio coding application, the entity considered is a compressed version of a reference (uncompressed) signal and the latter signal can simply be defined as the target. However, the case of spatial enhancement systems for headphone reproduction described later in this thesis differs in the sense that the optimal spatial reproduction over headphones to be employed as a target for the measurement of quality is actually not known in advance and should therefore be identified in the process of quality assessment. Secondly, it should be noted that different targets might co-exist for a given application, for example when differences between individual preferences are taken into account. In this case, only one of the targets is selected for the measure of quality.

2.2.2 Physical, sensory and affective characterizations of sound

The issue of quality assessment for any type of sound like noise, speech or music is always bound to the measurement of some characteristics of the entity under study but in practice such a characterization can be approached from very different perspectives, as illustrated in Figure 2.1. In a broad sense, two separate domains might be defined, considering that an acoustic event can be measured with either instrumental sensors or human sensors. The instrumental measurement of a physical sound signal can be performed with simple or advanced microphone techniques and a physical attribute profile can usually be derived from such a measurement using one or several pre-defined criteria. Examples of acoustical attributes include measures like sound pressure levels with e.g. A- or B-weighting, frequency content analysis, reverberation time or the speech intelligibility index defined in ANSI S3.5 (1997).

When human sensors are employed, the quantity of interest is the auditory sensation formed by the listener when exposed to the physical sound signal. However, an additional separation has to be defined in this domain because two distinct forms of measurements can be considered as illustrated in Figure 2.1. The sensory measurement of the auditory sensation on the one hand can be understood as the elicitation of objective responses to the properties of a stimulus, as perceived by the human senses (Piggott *et al.*, 1998). The affective measurement on the other hand implies a more subjective and global approach to the evaluation of the auditory sensation.

The distinction between the sensory and affective sides of the human measurement method has been widely exploited in the field of food science (Stone and Sidel, 1993),

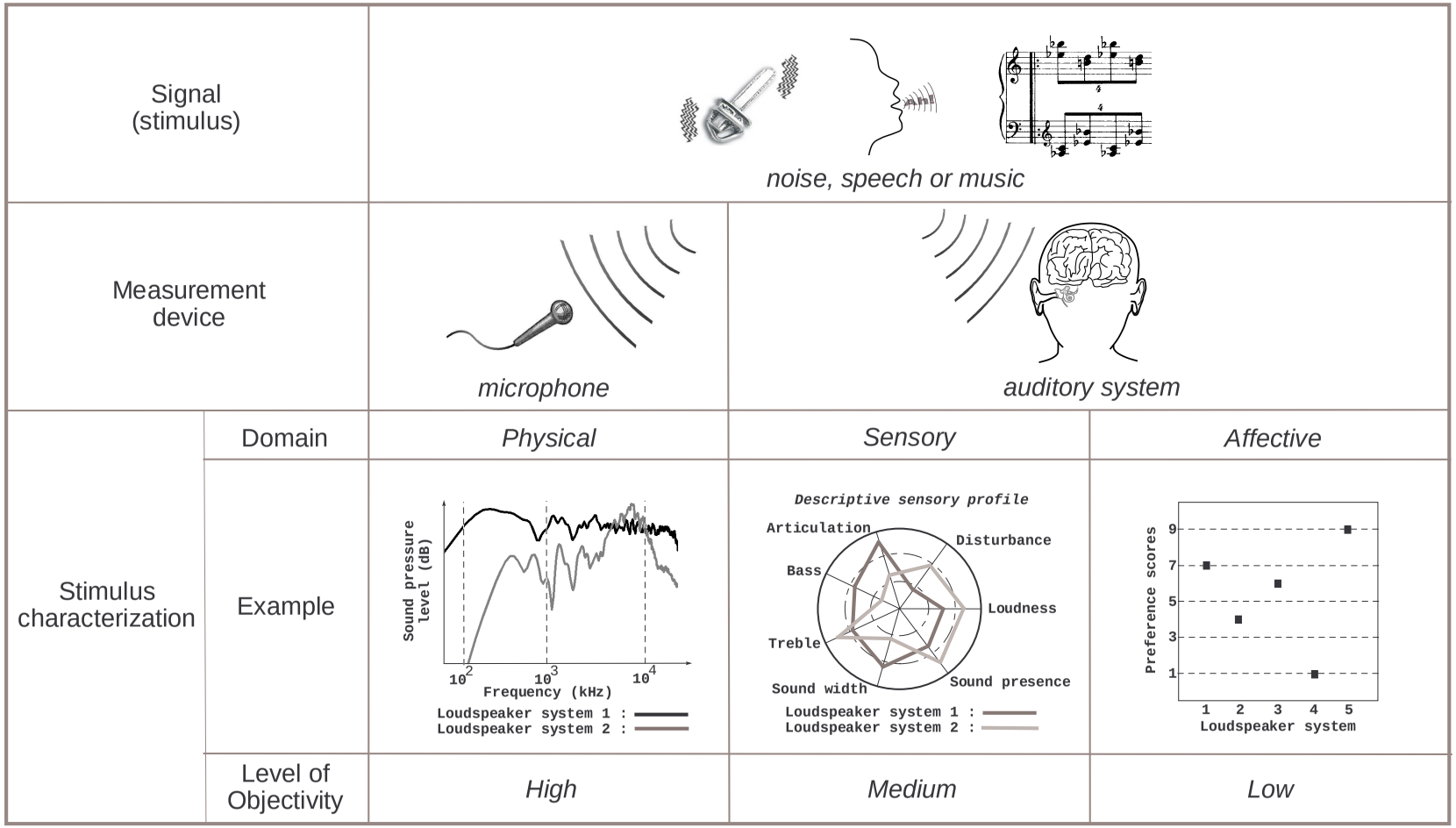


Figure 2.1: Classification of measurement methods illustrating the distinct nature of reproduced sound characterizations in the physical, sensory and affective domains.

which goal is usually to optimize consumers' preference for a product based on its perceived sensory characteristics. This approach forms the basis of the preference mapping method to be discussed later in this section. Similarly, Nunnally and Bernstein (1994) define two types of psychometric responses, namely *judgments* (relating to the sensory domain) and *sentiments* (relating to the affective domain) and they note that the correctness of an answer is only a valid concept in the former case. The separation between the sensory and affective domains has also been described in the field of audio by e.g. Pedersen and Fog (1998) or Bech and Zacharov (2006).

These three domains offer very different perspectives on the objects under study and the associated measurement techniques have their specific strengths and limitations. An important difference between these three approaches concerns the level of objectivity of the resulting sound characterization. To illustrate this idea, let's consider the measurement device in each of the three domains, as presented in Figure 2.1. In the physical case, one would be confident that a similar characterization is obtained from two different measuring instruments, for example the sound pressure level of a noise source measured with two calibrated microphones should not differ much. The sensory measurement follows the same principle and a sensory characterization made by two different subjects is expected to be similar to a certain extent. Characterizations made in these two domains can therefore both be considered as objective in this respect[2]. It should be noted that the level of objectivity of the measurement made by a human subject is usually lower than what can be achieved by an instrumental measurement. However, this issue is carefully addressed in sensory analysis methods by employing a group of trained subjects to insure an appropriate level of objectivity as will be discussed further in this chapter.

Considering now the affective domain, the assumption that two individuals would give a similar judgment about a perceived stimulus does not always hold. In the example shown in Figure 2.1, it is possible that substantially different preference scores would be obtained from two subjects and an example of such divergence between listeners in preference ratings of loudspeaker systems has been illustrated by Lorho (2006). Similar results have been reported in other fields of psychoacoustics, for example by Susini *et al.* (2004) who identified two sub-populations of listeners differing in their preference for sounds produced by air-conditioning systems. In that sense, the affective measurement made by a person is not expected to be highly objective because it usually involves aspects such as background, expectations, emotions, mood, etc. which can affect judgments in a subjective and individual way.

Coming back to the idea of correctness used by Nunnally and Bernstein (1994) to differentiate psychometric responses in the sensory and affective domains, it can be stated that while the criteria for correctness are imposed in a sensory measurement to achieve objectivity, such criteria are not forced on test subjects in an affective measurement but instead common patterns are identified from a large number of individual judgments and an objective criterion can be defined subsequently at a global level.

[2]In this context, the term 'objectivity' implies a focus on the object (stimulus) being measured by a given instrument and highlights the analytical nature and the intended reproducibility of the measurement being made on this object.

2.2.3 Relationships between sound characterizations and quality evaluation

The relationship between quality evaluation and the different sound characterizations presented above depends largely on the type of application but the variety of quality evaluation methods found in the field of audio follow similar underlying principles. A simple model of sound quality optimization problem can be utilized to illustrate how the different sound descriptions might be linked to perceived sound quality, as shown in Figure 2.2. The system under evaluation considered in this diagram is fairly generic and could represent any type of device producing sound. The output of this system constitutes the stimulus of interest and might either be an unwanted sound produced by the device directly, e.g. the noise generated by a ventilation system, or the transduced version of an input signal, e.g. the speech or the music played over a loudspeaker system. It should be noted that in this latter case the stimulus generated by the device reproducing the sound depends not only on the system under study but also on the input signal. The stimulus under test can therefore be seen as an interaction between these two elements, which adds some complexity to the problem of quality evaluation in practice, as will be further described later in this thesis. Figure 2.2 also illustrates the presence of some external parameters relating to the system under study, which constitute the experimental factors (independent variables) of interest and which represent usually the main elements of control in this problem of system quality optimization. In practice, experimental factors could be different noise attenuation strategies in the ventilation system example while in the loudspeaker reproduction example, this might be several transducer solutions or different settings for a signal processing algorithm.

On the measurement side, the stimulus or set of stimuli under study can be quantified in any of the physical, sensory or affective domains, leading for example to the characterizations presented in Figure 2.2, which illustrate approaches commonly encountered in the field of acoustics. Sound quality can usually be controlled by some external parameters relating to the system under study but the quality optimization process relies on some characterization of the system output, which can take different forms depending on the application of interest. Considering for example an over-simplified problem of noise attenuation, quality might be linked directly to the physical domain using an instrumental measure of an overall level of noise for example. In this case, the quality of the system could be optimized by adjusting the external parameters to reduce the noise level. However, if the overall level of noise cannot be attenuated in such a simple way, the goal of the quality optimization might then shift to reducing the perceived annoyance for example. As the application of a valid instrumental measure would be difficult in this scenario, quality needs to be approached from the affective domain with e.g. an overall acceptability measure. In practice, an optimal quality might be obtained by tuning the external parameters to achieve the lowest perceived annoyance, as illustrated in Figure 2.2. Finally, in more complex cases where instrumental measures made in the physical domain and overall measures made in the affective domain would fail to bring detailed enough information for the quality optimization process, a more accurate characterization produced in the sen-

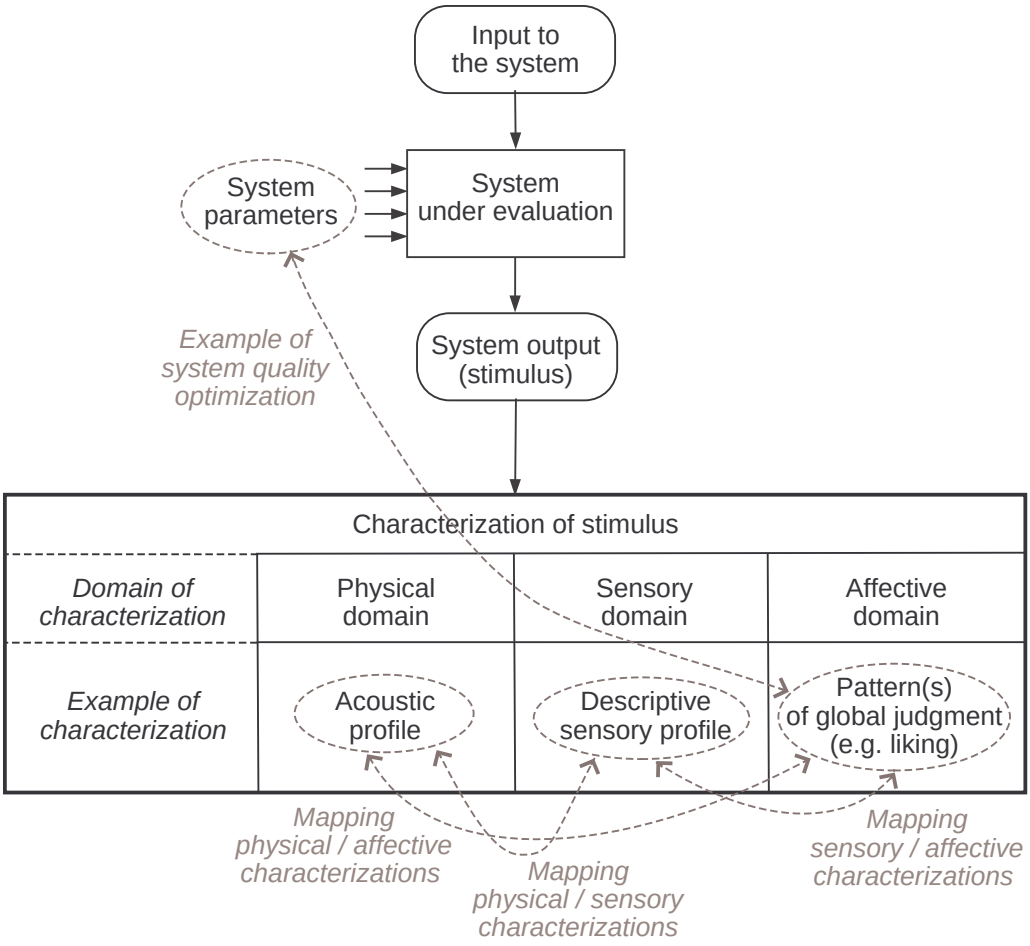


Figure 2.2: Relationships between sound event measures and sound quality evaluation.

sory domain can be considered. The application of this approach has been illustrated for example by Fastl (1997) with a noise quality optimization procedure based on the psychoacoustic (/sensory) measures of loudness, sharpness and fluctuation strength.

It should be noted that psychoacoustic measures like loudness or sharpness can in fact be derived directly from physical measurements. This shortcut is made possible by applying predictive models of these sensory characteristics to the recorded acoustic signals. Such models have been developed in the specific domain of loudness by e.g. Stevens (1961), Zwicker and Fastl (1990) and Moore *et al.* (1997) and an implementation of various psycho-acoustical algorithms can be found for instance in the PsySound3 software package of Cabrera *et al.* (2007). This approach illustrates an important topic of research concerned with the study of relationships between different sound characterizations. Combined measurements can be performed with the aim of building models to predict the measure(s) of one domain from the measure(s) of another domain. The possible mappings are illustrated in Figure 2.2 and many examples can be found in the literature for these different cases. For example in the context of loudspeaker sound quality, Toole (1986) studied the relationships between

listener preferences and physical measurements of loudspeakers; models to predict sensory attributes of spatial impression from acoustic (physical) measurements have been developed for auditory source width (Barron, 1971) and listener envelopment (Beranek, 1996); and Zacharov and Koivuniemi (2001b) developed a set of sensory attributes to describe spatial sound reproduction systems and created a predictive model of preference based on this sensory characterization.

2.2.4 A framework for perceived quality evaluation

The simplest form of quality optimization presented in the previous paragraph involving directly the physical domain are unfortunately rarely appropriate for systems producing relatively complex output, i.e., stimuli including several perceptual aspects (e.g. timbral and spatial aspects in 3D sound) or even multimodal aspects (e.g. audio-visual quality). In such cases, perceived quality can be approached from the affective domain through an overall evaluation of the perceived sensation (ITU-T P.800, 1996; ITU-R BS.1116-1, 1997; ITU-R BS.1534-1, 2003) but finding its direct relationship to instrumental measures is usually a difficult task. The evaluation of spatial enhancement algorithms for stereo music reproduction over headphones considered in this thesis illustrates a double challenge in that respect. Firstly, the characterization of stimuli in the physical domain is complicated in this case by the lack of instrumental metrics for spatial sound and secondly, defining the optimal quality of the enhancement in this type of audio application is not straightforward.

Following several listening experiments on stereo enhancement systems reported by Lorho *et al.* (2002), Lorho and Zacharov (2004) and Lorho (2005a) (see Chapter 3 for details), it became clear to the author that a quality evaluation based on a global measure of preference alone is not appropriate nor sufficient because of its lack of accuracy. To address these issues, a different approach was therefore adopted in which the sensory domain plays a central role in the definition and quantification of perceived quality. Considered on its own, the sensory characterization constitutes an objective description of the stimuli that relates better to the actual perception of listeners than common instrumental/physical characterizations. Furthermore, when the goal is to study relationships between different descriptions for complex perceptual problems, the use of a sensory characterization offers better perspectives to build accurate models in comparison to, e.g., a direct mapping between instrumental measures and overall affective measures. Preference mapping is an example of this type of approach combining two domains as described next.

Preference mapping belongs to the group of techniques to relate two independent sets of measurements made on the same objects. The objective of this combined analysis is usually to establish relationships between a set of affective judgments (e.g. preference or acceptability) and a set of characteristics (e.g. sensory or instrumental). Considering the example of a sensory profile produced by a panel of assessors and the preference judgments obtained from a large group of consumers for the same set of products, preference mapping can be employed to identify the most important sensory attributes driving preference while taking into account possible differences in judgment between consumers. This tool also offers different models to predict

the preference of new products based on existing sensory profiles and to identify the sensory characteristics of an ideal product for a given cluster of consumers.

Preference mapping techniques have been developed in the field of psychometry in the 1970s and have been widely used in sensory science (Greenhoff and MacFie, 1994; Schlich, 1995; McEwan, 1996). Carroll (1972) suggested the labels internal and external analysis, referring to distinct ways to handle the two data sets. While internal preference analysis is centered around consumer preferences, external preference mapping focuses on sensory information. In both cases, the primary data set is decomposed by multivariate analysis, e.g. principal component analysis (PCA), and the second data set is employed as a complementary source of information by means of regression analysis. The MDPREF method for internal analysis and the PREFMAP method for external analysis were originally developed by Chang and Carroll (1969) and Carroll (1972) respectively. Alternative methods have also been proposed later including the partial least square regression (PLS-R) introduced by Wold *et al.* (1983) which is a technique commonly applied to extract the sensory information that relates best to consumer preferences. More recently, new preference mapping techniques have been developed for sensory science applications by, e.g., Danzart *et al.* (2004) and Rivière *et al.* (2005).

Figure 2.3 illustrates how external preference mapping can be utilized in the process of quality assessment with the aim of developing an objective measure of perceived quality. Considering the model described in the previous paragraph (Figure 2.2), the aim is to measure and possibly optimize the perceived quality of the system under study based on a set of stimuli defined from the different external parameters of this system; this could be for example a stereo enhancement algorithm with different tuning parameters.

Two different characterizations need to be performed in this framework as illustrated in the upper part of Figure 2.3. The first characterization of the stimulus set is made in the affective domain. A global evaluation of likeness is considered in this example to identify subjects' overall impression of the different stimuli (note that the selection of the question depends on the goal of the quality optimization process). However, this affective characterization alone brings usually limited actionable information to the iterative process of quality optimization unless a complex experimental design is employed for the study as, for example, a response surface methodology on the external parameters (Myers and Montgomery, 2002). Therefore, a second characterization of the same stimulus set is made in the sensory domain through the descriptive sensory profiling approach.

The two resulting characterizations are complementary and can be exploited in the next step of the process where perceived quality is modeled and optimized. The external preference mapping technique is employed to relate the descriptive sensory profiling data to the global likeness data derived from this stimulus set. Through this analysis technique, a mapping model can be built between these two data sets, the most important sensory attributes responsible for liking or disliking can be identified and one or several optimal liking regions can be predicted. An 'optimal point' represents a target for the quality optimization process and it can be specified in the affective domain as a point of optimal likeness but also in the sensory domain as a

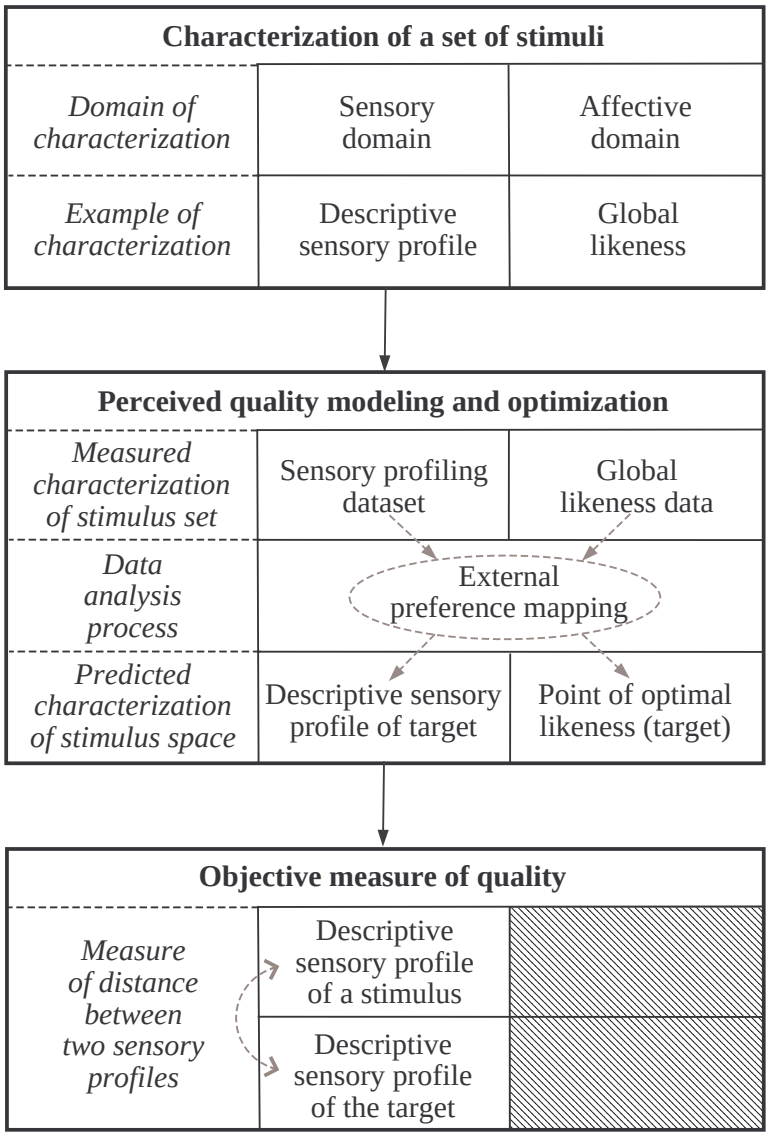


Figure 2.3: Application of the preference mapping technique in the context of perceived quality optimization.

target descriptive sensory profile, which is illustrated in the middle part of Figure 2.3.

Finally, recalling that quality was defined earlier in this section as a measure of the distance between the character of an entity and the character of a target, an objective measure of perceived quality can be derived by comparing the descriptive sensory profile of a (measured) stimulus to the descriptive sensory profile of the (predicted) target[3].

[3]It should be noted that this preference mapping framework can also be applied between the physical and affective domains, for example by replacing sensory attributes by acoustical measures on the left side of Figure 2.3.

2.3 Sensory analysis

Applying the preference mapping approach described above in the context of perceived quality optimization requires a careful consideration of all the stages in the process. This includes an appropriate selection of stimuli to cover the whole perceptual space under study but also the choice and application of a suitable methodology for both sensory and affective testing and finally the development of a valid predictive model from the two data sets gathered. As the application of sensory analysis to the perceived quality evaluation of sound reproduction is the main focus of the research effort described in this thesis, a more thorough presentation of this measurement technique is given next. The topic of assessor classification for sensory testing is addressed first. Then, an overview of sensory analysis techniques highlighting the commonalities and specificities of the different strategies found in this type of testing is provided. Finally, more details on descriptive sensory analysis methods using verbal elicitation are given and the application of this family of techniques to perceived sound quality evaluation in the context of this thesis is justified.

2.3.1 Assessors and panels for sensory testing

Sensory analysis was presented in the previous section as one of the two approaches to characterize stimuli by means of human sensors and was defined as a method to elicit objective responses to the properties of a stimulus, as perceived by the human senses. The author discussed the difference in goals between sensory and affective testing and highlighted the fact that objectivity is intended in the former case while subjectivity is taken into account in the latter case. The distinction between these two forms of testing has a significant influence on the type of subjects to be employed for the task. Scriven (2005) discusses the two stages of human response to stimuli as a primary response resulting from the descriptive process of recognizing and measuring the stimulus and a secondary response resulting from the affective process of forming a judgment about what is perceived, e.g. liking or acceptability judgements. Scriven (2005) argues that typically, naïve subjects are only conscious of the secondary response but will develop their awareness of the primary response through training. Lawless and Heyman (1998) make also this distinction between naïve and trained subjects and warn about matching test methods and respondents. It is usually viewed that obtaining affective measures such as liking, preference, or acceptance is best achieved with naïve subjects.

The topic of subject categorization has been addressed in a number of audio standards and recommendations including CCITT (1992); ITU-T P.800 (1996); ITU-R BS.1116-1 (1997); ITU-T P.831 (1998); ITU-T P.832 (2000). However, the terms 'untrained', 'naïve', 'experienced' and 'expert' to define subjects participating in audio testing do not appear to be employed very consistently in the literature. In comparison, the classification of test subjects developed in the field of food science reflects better the distinction between affective and sensory testing and has therefore been adopted in this thesis. The categorization of test subjects employed in sensory analysis and the processes applied to develop the sensory expertise of assessors have been formulated in the ISO standards 8586-1 (1993) and 8586-2 (1994). The proposed ter-

Assessor type	Definition
Assessor	Any person 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, ageing, etc.

Table 2.1: Definition of assessor types employed in sensory analysis according to the ISO standards 8586-1 (1993) and 8586-2 (1994).

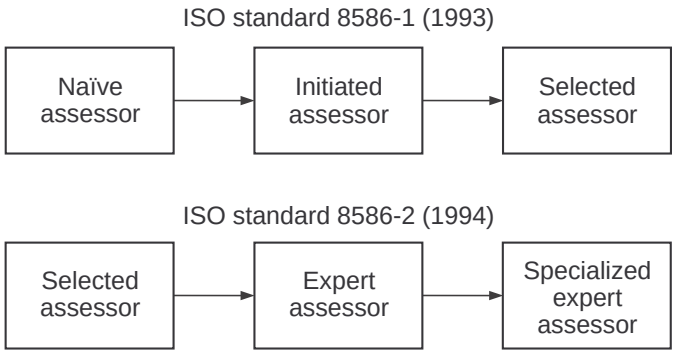


Figure 2.4: Summary of the different stages in the development process of sensory assessors according to the ISO standards 8586-1 (1993) and 8586-2 (1994).

minology reproduced in Table 2.1 covers various types of assessors. The distinction made in this categorization system between the terms 'naïve', 'selected' and 'expert' highlights clearly the meaning of sensory expertise, that is, the ability to carry out a sensory test with skills (sensory acuity, reliability, etc.) validated through a formal screening and training process. These two standards outline also the stages in the development process of sensory assessors as summarized in Figure 2.4 and describe the

procedures applied in food science applications to screen, train and monitor selected assessors (ISO 8586-1, 1993) and (specialized) expert assessors (ISO 8586-2, 1994).

The importance of subject skills has also been acknowledged and studied in the domain of audio perception by e.g. Bech (1992) and different procedures have been proposed in the literature for listener pre-selection (Mattila and Zacharov, 2001; Isherwood *et al.*, 2003; Wickelmaier and Choisel, 2005) and training of listeners for timbral perception (Quesnel and Woszczyk, 1994; Quesnel, 1996, 2002) or spatial perception (Neher *et al.*, 2002).

Coming back to the issue of objectivity in measurements made in the physical and sensory domains, an additional specificity of sensory analysis should be highlighted. While a single measuring instrument is usually employed in the physical domain (for example one microphone), a measurement made in the sensory domain requires a group of human subjects often referred to as a 'sensory panel'. In this respect, the measuring instrument of any sensory analysis method should be understood as the group of individuals as a whole. A panel of assessors is employed for sensory testing rather than just one person to take into account the fact that human subjects are not equally sensitive to sensory stimuli. Individuals might also vary in their ability to discriminate different perceptual aspects of the stimuli and can be subject to judgment bias. Therefore, the use of a panel of assessors yields a more stable and reliable description of the stimuli, which help ensuring an appropriate level of objectivity for the intended sensory measurement. The drawback of this special type of measuring instrument is that it requires additional efforts in terms of panel handling as will be illustrated later in this thesis.

A last remark on sensory panels can be brought up to highlight an additional distinction between sensory analysis techniques based on the intended goal. The term 'consumer panel' is sometimes found in the literature (see for example Jack and Piggott, 1991) implying that sensory testing is made on assessors that do not fit the requirements of expertise described above. In such cases, no special care is taken to screen or train subjects and testing can sometimes be performed outside of the laboratory, e.g. in real home use conditions for food products. Similarly, O'Mahony (1995) makes a distinction between two types of sensory evaluation. The first type called 'Analytical Sensory Testing' emphasizes the reliability and sensitivity of the measurement and employs therefore highly trained assessors, which can not be considered as representative of consumers. The second type called 'Measurement of Consumer Perception' employs assessors selected to be representative of a target consumer group and who are usually given less training time[4]. O'Mahony's paper describes also how different sensory analysis methods might fall into either the first or the second category and similarly Lawless and Heyman (1998, page 16) considers that "*every sensory test falls somewhere along a continuum where reliability versus real-life extrapolation are in a potential trade-off relationship*". This additional distinction

[4]Note that both of these panel types are employed for the purpose of sensory analysis. On this matter, O'Mahony (1995) warns not to mix the two categories 'Measurement of Consumer Perception' and 'Affective Testing'.

between sensory methods illustrates a more contrasted but also a more realistic view of sensory testing methods and sensory panels. In practice, applications of sensory testing with panels of either experts or naïve subjects are found in the literature, the latter being sometimes of considerable size, e.g. 150 consumers in a perceptual free sorting study reported by Faye *et al.* (2004). It is therefore important when designing a perceptual experiment to define precisely the goal of the intended measurement and make the appropriate selection of both the sensory testing method and the associated panel of assessors.

2.3.2 An overview of sensory analysis methods

Figure 2.5 presents an overview of the sensory analysis methods commonly encountered in the field of sensory science. The main families and sub-families of sensory tests are illustrated in this classification and a non-exhaustive set of examples is provided for each of these groups of techniques. Sensory analysis methods can be loosely separated into two groups: discriminant methods and descriptive methods. The former group addresses the issue of measuring perceived differences between stimuli, whereas the latter group aims at the identification, description and quantification of the sensory attributes of stimuli by trained human subjects (Piggott *et al.*, 1998).

Discrimination testing is usually applied to study small differences between stimuli, e.g. to measure the perceived degradation introduced by a high-quality audio compression algorithm. Discriminative tests include various methods such as the triangle test (ISO 4120, 2004) and the (binary) paired comparison (ISO 5495, 2005). These two examples illustrate respectively the global approach and the specific approach to difference testing. In the triangle test, assessors are presented with three stimuli, two of them being similar, and are asked to identify the stimulus the most different from the two others. Assessors are therefore not required to focus on a specific aspect of the stimuli. On the contrary, the aim of the paired comparison is to identify which of two stimuli has the most of a designated perceptual characteristic, e.g. loudness or source width in the field of audio. Statistical handling of data resulting from discrimination tests relies usually on Thurstonian modeling (Thurstone, 1927) and numerous applications to sensory evaluation have been described in the literature (see for example O'Mahony and Rousseau, 2003).

Descriptive analysis forms the second group of sensory methods. It can be seen as a more exploratory approach and is often described as the most sophisticated tool in sensory science (Lawless and Heyman, 1998). A large number of techniques have been specifically designed for this purpose, mainly in the domain of food science but also in other perceptual domains. Bech and Zacharov (2006) and Neher *et al.* (2006) reviewed these methods and their application to the field of audio. Different approaches to the problem of sensory attribute elicitation can be considered as illustrated in Figure 2.5. This includes a) the methods relying on a verbal description of perceived sensations, e.g. techniques employing a vocabulary development process with a group of assessors; b) the non-verbal elicitation methods, i.e. techniques based on body gestures and c) the methods working without direct sensation labeling.

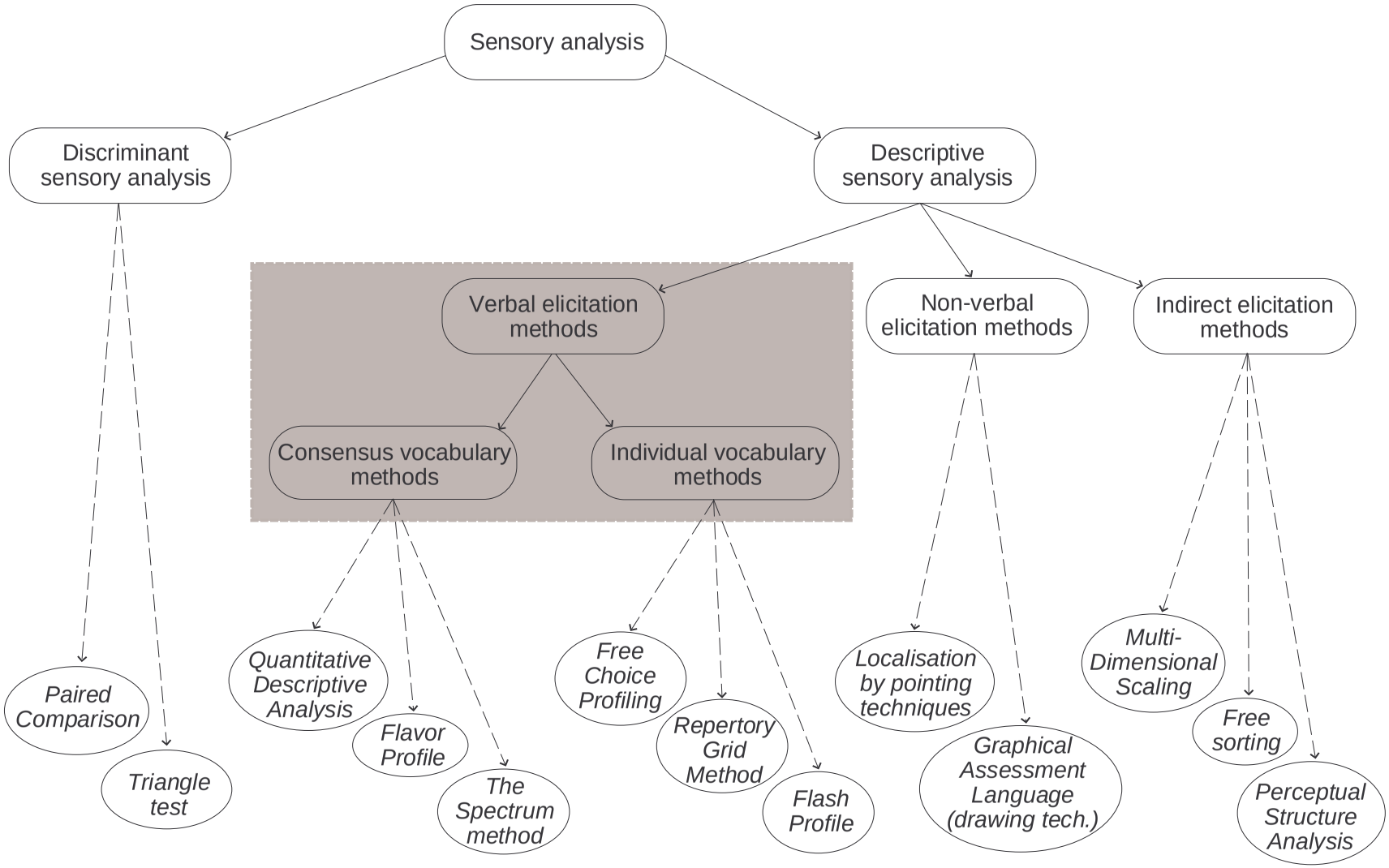


Figure 2.5: Overview of the sensory analysis methods commonly encountered in the field of sensory science illustrating the two principal families and their sub-families with some examples.

Techniques based on verbal elicitation form the largest group of descriptive analysis methods and have been developed and extensively utilized in the field of sensory science. Most of these methods are based on the same principle of sensory characterization using quantitative scales associated to verbal descriptors. However, two distinct routes exist for establishing the sensory descriptors as illustrated in Figure 2.5. On the one hand, **consensus vocabulary** (CV) methods use a panel of assessors to develop a common set of descriptors characterizing the sensory properties of the stimuli under investigation. Examples includes the Flavour Profile method (Cairncross and Sjöström, 1950), Quantitative Descriptive Analysis (Stone *et al.*, 1974) and the Spectrum method (Meilgaard *et al.*, 1991). On the other hand, **individual vocabulary** (IV) methods let each assessor develop his or her own set of sensory descriptors. Examples includes Free-choice profiling (Williams and Langron, 1984), the Repertory grid technique (Kelly, 1955) and Flash profile (Delarue and Sieffermann, 2004). More details about these two approaches to verbal elicitation will be given in the latter part of this section and in the next chapters.

Techniques based on non-verbal elicitation form a second group of descriptive analysis methods also aiming at a direct elicitation of perceived sensations but without using a formal set of verbal descriptors. This approach has been employed mainly in the field of audio research. Mason *et al.* (2001) reviewed the different forms of non-verbal elicitation and discussed their advantages and limitations for spatial sound perception. Several techniques based on body gestures have found interesting applications in this context, the rational being that verbal elicitation is not always appropriate to describe the complexity of an auditory space. For example, pointing techniques have been commonly used in sound localization experiments in which assessors are asked to indicate the direction of the auditory event they hear. Methods found in the literature include head movement (Chung *et al.*, 2000), hand pointing with a tracked device (Gröhn *et al.*, 2002), laser pointing (Choisel, 2003) or direction mapping through an external device (Riederer, 2005). Drawing techniques have also been exploited for more advanced elicitation experiments involving spatial sound reproduction systems. Martens (1999) employed this approach to evaluate the spatial extent, shape, and location of the auditory spatial image of different loudspeaker setups. Ford *et al.* (2002) developed a graphical assessment language (GAL) for automotive multichannel sound system evaluation including the two important spatial attributes of 'image skew' and 'ensemble width'. Usher and Woszczyk (2003) employed a similar approach to study listeners perception of different loudspeaker configurations.

Techniques based on indirect elicitation form the third group of descriptive sensory analysis methods illustrated in Figure 2.5. The test methods included in this group differ notably from the two approaches already discussed in the sense that they do not require the subjects to elicit directly the perceived sensory characteristics of the stimuli.

Multidimensional scaling (MDS, Carroll, 1972) is an example of indirect elicitation method commonly found in sensory analysis. Applications in the field of audio include e.g. Miller and Carterette (1975), McAdams *et al.* (1995) and Choisel and Wickelmaier (2005). In this technique, only perceived similarities or dissimilarities

between stimuli are collected from the assessors. A distance matrix is created from the gradings obtained for a combination of stimulus pairs and a spatial map of the perceived differences between the stimuli can be built using different analysis models such as classical MDS or weighted MDS (Popper and Heymann, 1996). It should be mentioned however that distance matrices alone do not offer a way to interpret directly the perceptual dimensions associated with the spatial map because no labeling of the sensation is asked from the subjects.

Free sorting is an alternative indirect elicitation approach based on a similar principle of perceived similarity evaluation. This method requires assessors to create groups containing stimuli that are perceived as similar according to their own criteria but in addition, assessors can be asked after the sorting task to describe each group of stimuli with verbal descriptors. This a posteriori labeling facilitates the interpretation of perceptual dimensions at the analysis step. Recent applications of this technique can be found for example in the studies of Faye *et al.* (2004) and Cartier *et al.* (2006) or in the Interpretation-based Quality (IBQ) approach developed by Nyman *et al.* (2006). A similar sorting approach referred to as projective mapping has also been used by e.g. Risvik *et al.* (1994) and Perrin *et al.* (2008) in which subjects translate their perception of stimulus differences directly into distances in a two-dimensional space, for example using a sheet of paper. All these perceptual grouping techniques seem to be well adapted to produce a coarse perceptual map that can be interpreted semantically when a verbal description task complements the indirect elicitation.

Perceptual Structure Analysis (PSA) is another example of indirect elicitation technique recently introduced in the field of audio by Choisel and Wickelmaier (2005) and Wickelmaier and Ellermeier (2007). This approach inspired from Heller's theory of semantic structures (Heller, 2000) is based on the idea of separating the processes of identification and labeling of perceived characteristics. In practice, subjects are presented with three stimuli and are simply asked to indicate if the first two stimuli share a common 'feature' with the third stimulus or not. Upon careful verification that subjects use consistently the features they perceived in the stimulus set, a representation of the individual perceptual structure can be derived indirectly. PSA has been successfully applied by Choisel and Wickelmaier (2007) to develop a set of auditory attributes describing the perception of multichannel sound reproduction.

Definition of the terms 'attribute', 'descriptor' and 'dimension'

The terminology found in the literature to describe aspects relating to perception include many terms such as 'attribute', 'feature', 'sensory concept' or 'construct', 'sensory descriptor', perceptual 'dimension' or 'component', etc. The classification of sensory analysis methods presented above illustrates the wide range of approaches available to elicit sensory responses from human subjects and it also highlights some major differences between these techniques in terms of expected outcome. Based on this review, the author proposes to group the terms commonly encountered in the field of sensory science in three classes which are defined next. The usage of this terminology in the present thesis is also clarified below.

The term '**attribute**' is commonly found in the context of sensory analysis and

is usually employed to describe a perceptible characteristic of an object. This term focuses more on the mental representation of a perceived entity than on the label associated with the perceived sensation. The term '(auditory) feature' found in the Perceptual Structure Analysis method carries a similar meaning[5] and the fact that this indirect elicitation method makes a clear separation between the processes of identification and labeling of perceived characteristics illustrates well the idea that attributes can be conceived without a descriptive word association. Two other terms encountered in the field of sensory science relate closely to the formation of conceptual entities, namely the term 'sensory concept' as defined by e.g. O'Mahony (1991) to be discussed in Chapter 4, and the term '(personal) construct' as defined by e.g. Kelly (1955) to be discussed in Chapter 7.

The term '**descriptor**' relates also to a perceived characteristic of an object but with a tight association to the label given to this sensation. A 'sensory descriptor' is mainly employed in verbal descriptive analysis and is usually specified with a clear definition. Giboreau *et al.* (2007) highlighted the importance of defining accurately descriptors in consensus descriptive analysis and they proposed guidelines to improve the structure and content of definitions. A set of sensory descriptors elicited by a panel of assessors for a given set of stimuli forms a (descriptive) 'vocabulary'. The terms 'glossary', 'lexicon' and 'language' are also encountered in the literature with a similar meaning. Note however that the latter term refers formally to a system comprising of both the symbols and rules for symbol manipulation, which might be argued to go beyond the scope of just a group of verbal descriptors. Finally, it should be mentioned that the terms 'attribute' and 'descriptor' are often used without a clear distinction in the context of verbal descriptive analysis. This is mainly due to the fact that a sensory concept is always tightly associated to a label and a verbal description in this type of sensory analysis.

The term '**dimension**' is usually employed to describe the underlying structure of a given stimulus set. Unlike an attribute or a descriptor which can be both considered as a measurable perceived characteristic, a dimension is not a direct observable[6]. A multivariate analysis technique has to be applied to extract these dimensions also called '(principal) components' or 'factors', each of them being expressed as, or approximated by, a linear combination of manifest variables. The resulting perceptual dimensions can be interpreted from the manifest variables, which is a process known as reification. It should be highlighted however that these dimensions only reflect the statistical variation in the multivariate data describing the set of stimuli under study.

The term 'perceptual dimension' can convey this meaning but it is also employed to describe an essential characteristic of a representative set of stimuli identified and interpreted from one or several experiments. It appears that the statistical restrictions of the above definition of 'dimension' are not always valid in this broader context. Therefore, to make a clearer distinction between these two meanings, the author will employ

[5]The term 'attribute' can also be found in Heller's theory of semantic structures (Heller, 2000).

[6]These two different types of data are usually referred to as a 'manifest' and 'latent' variables respectively.

the term 'perceptual direction' (in a latent space defined by several dimensions) to describe more loosely a sensory pattern identified in one or several multivariate data analyses and associated to a group of co-varying manifest variables or even a single variable.

These three generic terms will be encountered within the present thesis in which two descriptive analysis experiments using a verbal elicitation approach will be reported. The two terms 'sensory attribute' and 'sensory descriptor' will be used in the context of the CV of headphone sound reproduction presented in Chapters 4 and 5 but the term 'dimension' will also appear whenever multivariate analysis methods are applied to a given set of stimuli. The second experiment to be reported in Chapter 9 employed an IV approach and produced sensory attributes and descriptors which are less well defined at a panel level than in the case of a CV as will be discussed in Chapters 7 and 9. For this reason, the concept of latent space and dimension is essential in this type of experiment and will be largely exploited. The analysis of individual vocabulary profiling data will also make use of the concept of 'perceptual direction' for the purpose of result interpretation.

2.3.3 Descriptive sensory analysis using verbal elicitation

Selection of a verbal elicitation approach

The overview of sensory analysis presented above illustrates the abundance of techniques available for this type of measurement. The selection of an appropriate technique is not always an easy task in practice and it can sometimes be a topic of debate amongst sensory scientists. However, beyond preferences for a certain class of methods or familiarity with a given sensory analysis technique, rational justifications for methodological choices are always desired considering especially the large effort generally invested in the implementation of such sensory testing methods.

The perceptual study reported in this thesis relates to headphone sound perception and includes both an exploratory and an application-specific investigation. One aim of this work was to study the perception of sound reproduced over headphones in a broad sense while the other aim was to evaluate the perceived quality of a given set of stereo enhancement systems for headphone reproduction. The use of a verbal elicitation approach for this study was inspired by the two consensus vocabulary profiling experiments (Mattila, 2001b; Zacharov and Koivuniemi, 2001b) performed earlier at the Nokia Research Center by the author's colleagues. These research works illustrated that verbalization was an appropriate approach to elicit and communicate perceived impressions despite the complexity of the sound stimulus sets investigated in these two studies. Verbal descriptive analysis has also the advantage of being well documented in the literature with at least three text books covering thoroughly the topic (Meilgaard *et al.*, 1991; Stone and Sidel, 1993; Lawless and Heyman, 1998) and several standard recommendations in the field of food science (ISO 11035, 1994; ISO 13299, 2003; Majou *et al.*, 2001). In addition, sensory analysis using verbal elicitation includes a wide range of applications to food products and non-food products (e.g.

image quality, tactile aspects, thermal comfort, etc.) reported in the literature, which offered an excellent source of guidance for the present study despite the difference in stimulus modality.

Some new experimentations with the class of non-verbal elicitation techniques were initiated in the field of audio during the process of the experimental work reported in this thesis (2003-2004), especially in the direction of drawing techniques (GAL) for spatial sound evaluation (Ford *et al.*, 2002). In the context of the present study however, both spatial and timbral aspects were considered important and drawing techniques designed to study mainly spatial characteristics of reproduced sound would therefore not be optimal. On this matter, it can be noted that most of the verbal elicitation techniques listed in the previous section do not make assumptions on the sensory aspects covered in the descriptive analysis but aim rather at an exploration of all the sensory aspects perceived in the stimulus set.

On the side of indirect elicitation methods, alternative techniques were also available. MDS has been commonly used in the field of audio perception but its application would have been problematic in this study because the direct collection of dissimilarities on n products requires the comparison of $n(n\text{-}1)/2$ pairs, which would yield a too large test design for the 50 stimuli included in first phase of this study (see chapter 5). Free sorting methods were less familiar to the author at this early stage of the research and were therefore not considered due to the (apparent) complexity of collecting both dissimilarity and descriptive data. However, in light of the recent publications (e.g. Faye *et al.* (2004) and the other references mentioned earlier), the potential of this approach should be acknowledged. Finally, it can be noted that the alternative approach of PSA as an indirect elicitation method was not developed yet during the present work.

As the two experiments described in this thesis employed a verbal elicitation approach, an overview of this type of descriptive sensory analysis is presented in the remaining part of this section to highlight its principle and the differences existing between the IV and CV development processes.

Considerations on the verbal elicitation process

Psychology is certainly the research field to survey when attempting to trace back the origins of current verbal descriptive analysis practices. The introspective structuralism for example might be considered as a precursor of modern verbal elicitation techniques. The idea of this psychological methodology developed during the 19^{th} century by Edward Bradford Titchener was to ask trained subjects to decompose their conscious experience into its fundamental elements by an analytical introspection. Through this type of introspective verbalization, it was possible to uncover elements of conscious experience such as sensations or feelings and to investigate the relationships between these elements. Current verbal elicitation techniques shares also some ideas with two more recent methodologies introduced during the 1950's in the field of psychology, namely the semantic differential developed by Osgood (1952, 1967) and the Repertory Grid developed by Kelly (1955).

The semantic differential was introduced by Osgood (1952) as a general method to measure meaning and can be described as a combination of controlled association and scaling procedures. In practice, a concept[7] and a set of bipolar adjectival scales are presented to a person. For each item (pairing of a concept with a scale) the task of the subject is to indicate the direction of his association and its intensity on a seven-step scale. Osgood's assumption was that linguistic encoding of meanings (abstractions) requires a carefully devised sample of alternative verbal descriptions that should be representative of the major way the stimuli and their associated meanings can vary.

Kelly's Personal Construct Theory (1955) is based on the idea that the construction system used by a person to perceive, understand, predict and control the surrounding world is composed of a finite number of dichotomous constructs. Kelly developed a technique originally referred to as the 'role construct repertory test' to explore such constructs, with a specific application to the measurement of personality in the field of psychology. This approach encourages introspection in a structured way by requiring subjects to define their own constructs by indicating in which way two out of three elements (usually persons in this psychology context) are similar to each other and different from a third one.

Despite the differences in both the underlying theory and the objects of investigation, it is easy to recognize elements of these psychological methodologies in current verbal descriptive analysis methods, namely the use of introspective verbalization, the analytical decomposition of mental experiences, the creation of verbal descriptors and the use of quantitative scales. One interesting aspect to note based on this brief historical background concerns the apparent assumptions made in descriptive sensory techniques. For example, there seems to be little theoretical justifications of the assumption that (trained) subjects can break down their perception into its constituting elements and can give a meaningful verbal description of these perceptual components. On this topic, the choice of a psychophysical model assuming that a subject can attend independently to perceptually separable features has been questioned for complex stimuli such as odors by e.g. Lawless (1999). Nevertheless, the validity of this approach is usually accepted empirically considering the successful applications in research and development over few decades for very diverse types of stimuli involving different sensory modalities. Another validity issue discussed in the literature relates to the ability or the inability of language to describe perception (see e.g. Samoylenko *et al.* (1996) for a review), the latter being sometimes a justification for the use of non-verbal sensory analysis techniques. It should be noted however that the issues relating to this topic have been discussed in depth by Civille and Lawless (1986) who outlined useful principles for a precise and reliable use of words to describe perception and who highlighted also the benefits of verbal elicitation.

To further illustrate the concept of verbal descriptive analysis, it is useful at this point to elaborate a bit on the processes experienced by a sensory assessor during

[7]Osgood uses the term 'concept' in a general sense to refer to the 'stimulus' to which a subject responds. For example, the meaning of abstract stimuli such as words can be investigated, as reported in Osgood (1967).

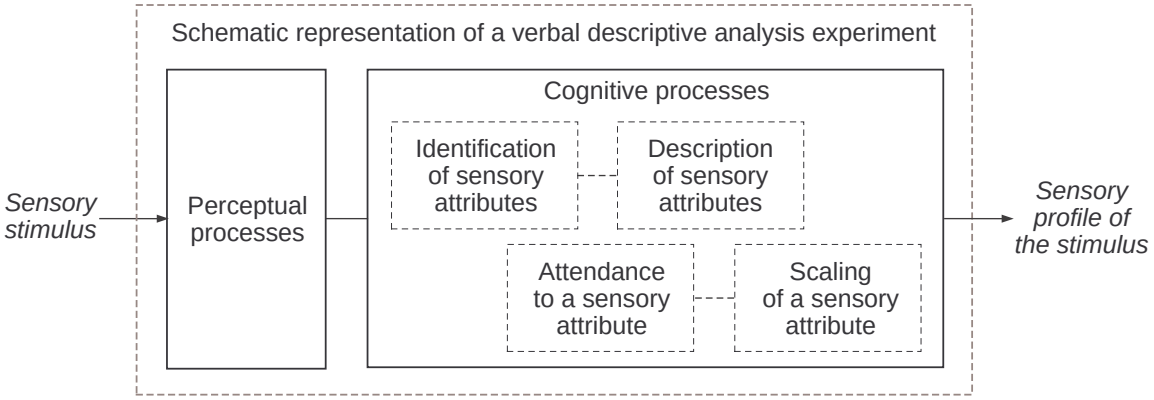


Figure 2.6: Schematic representation of the processes experienced by a sensory assessor during a verbal descriptive analysis experiment.

his/her assignment. The stimuli considered in this type of experiment are usually of complex nature, meaning that they cover multiple perceptual aspects possibly from several modalities and the task of the assessor is to identify, describe and quantify the sensory attributes of these stimuli. Figure 2.6[8] shows a simple representation of the processes experienced by an assessor during a verbal descriptive analysis experiment. This model includes two separate blocks corresponding respectively to the perceptual and cognitive processes involved in this task. When a stimulus is presented to the sensory assessor, a physical information (e.g. in the acoustic form) is transferred to the sensory domain through a series of complex perceptual processes, which results in the formation of a (complete) sensory event in the mind of the assessor. Then, a series of cognitive steps is required for the subsequent conversion of this sensory event into a quantitative description of its perceptual elements by the assessor. Different groups of cognitive processes can take place at different steps of a descriptive sensory experiment. For example, the identification and verbal description of the perceived attributes at the early stage of the experiment or the attendance to and the scaling of a given attribute at the final stage of the experiment. It is important to note in the diagram presented in Figure 2.6 that causal relationships between the perceptual and cognitive processes are not clearly stated because their interactions are not fully understood, as described for example by Blauert (1997) for the auditory domain. Similarly, the two sets of boxes included in the group of cognitive processes are drawn with dashed lines to illustrate the fact that these processes can not be considered strictly in isolation.

Consensus versus individual elicitation

The above description of the processes experienced by an assessor during a verbal descriptive analysis focused on a single sensory assessor but the fact that a panel of assessors is employed in sensory analysis rather than just one person has some impact

[8] A related description of a subject in an auditory experiment has been presented by Blauert (1997) and Mason *et al.* (2001).

on the application of this model. The way the sensory panel works in the two types of vocabulary development procedures earlier referred to as CV and IV methods (Section 2.3.2) differs considerably in terms of elicitation process and this can be reflected in the basic model presented in Figure 2.6 as follows.

In the case of IV methods, no external constraint is imposed during the elicitation process and assessors are free to identify, describe and use the sensory attributes as fits best their own perception. The model of a single assessor presented in Figure 2.6 applies therefore directly to this scenario. The situation is very different in the case of CV methods as the panel of assessors is required to agree on the meaning of the sensory attributes under consideration. The interaction between the assessors during the elicitation imposes large constraints on each individual. This aspect specific to CV methods can be represented in the diagram shown in Figure 2.6 as an additional external input in the form of verbal information, which modifies significantly the cognitive processes taking place in each assessor.

These fundamental differences between the two approaches have a large impact on the implementation of the associated sensory experiments and they also influence the analysis and interpretation of results by the experimenter as will be developed further in this thesis. A more thorough presentation of the IV and CV methods is left to Chapters 4 and 7 respectively and a detailed comparison of these two approaches is presented in Chapter 10.

2.4 Summary

This chapter introduced the topic of perceived audio quality with a discussion on concepts relating to quality. A definition of the term 'Quality' was proposed as a measure of the distance between the character of an entity and the character of a target. A framework for an objective evaluation of perceived quality was also presented based on the principle of preference mapping. In the second part of this chapter, an overview of sensory analysis was provided covering three aspects. Firstly, a clarification on the categorization of sensory assessors and sensory panels was provided. Secondly, a classification of descriptive sensory analysis techniques was proposed in which the specificities of three large families of methods were highlighted with a list of important examples from the literature. Thirdly, the sub-family of verbal descriptive sensory analysis techniques was reviewed and some important differences between the two approaches for verbal elicitation referred to as CV and IV development were highlighted.

Chapter 3

Reproduction and Perception of Spatial Sound Over Headphones

3.1 Introduction

Spatial sound is usually associated to multichannel loudspeaker systems but headphones are also employed for such applications. In this chapter, an overview of the techniques commonly applied for spatial sound recording/reproduction is provided and their related perceptual aspects are discussed. A literature review of spatial sound studies is presented and the topic of headphone sound reproduction and perception is addressed. Finally, a set of experiments on spatial enhancement techniques for headphone reproduction is briefly reported as an introduction to the perceptual studies relating to the same audio application to be presented in the following chapters.

3.2 Spatial sound reproduction

3.2.1 Spatial sound recording/reproduction techniques

The basic idea of spatial sound recording and reproduction can be loosely interpreted as the coupled problem of capturing a real acoustic scene, e.g. a musical performance at a concert hall venue, and recreating it virtually in an other space. An intuitive goal for this recording/reproduction process might be defined as a transparent reproduction of the original sound characteristics and this idea seems to be supported by the common use of concepts relating to fidelity, authenticity or naturalness in the literature on reproduced sound. Spatial characteristics play an important role in this process and can be affected by many technical factors both at the recording side and the reproduction side of the chain. In practice, the recording/reproduction process can include one, two or more than two channels depending on the technique considered.

The one-channel case offers a limited potential in terms of spatial characteristics in comparison to the 2-channel and multichannel approaches although monophonic audio signals do provide some level of spatial cues, such as a sense of distance and depth. Sound recording/reproduction techniques using two channels evolved throughout the

1900s and they include mainly two-channel stereophony and binaural stereophony[1]. The development of multichannel stereophony[2] started in the 1950's for cinema sound and followed with home cinema applications as early as the 1970's. To highlight the historical evolution of surround technology, a differentiation can be made between the older analogue matrix encoded formats and the more recent digital discrete multichannel formats. A short presentation of these different approaches to sound recording/reproduction is presented next (for a more complete review on this topic, see e.g. Rumsey, 2001).

The concept of stereophonic transmission first appeared in 1881 with the creation, patenting and demonstration of the Theatrophone by Clément Ader at the International Electrical Exhibition in Paris (Du Moncel, 1887). This system comprised a number of microphones, the signals of which were transmitted to remote sites employing two telephone lines. Reproduction occurred via two telephone earpieces, one placed on each ear, and was perhaps the first demonstration of the stereo headphone concept. Stereophony evolved throughout the 1930's with Bell Telephone Labs research on music transmission using multichannel systems, i.e., two or three microphones and loudspeakers (Snow, 1934). During the same period, binaural stereo[3] was patented by Blumlein with the development of the binaural microphone and associated recording techniques, as reported in the patent 394,325 (Blumlein, 1931). The commercial adoption of two-channel stereophony occurred in the 1950's and this format is still predominant in mainstream consumer audio applications. Current two-channel stereo techniques include different types of transducer setups at the recording side with, e.g. coincident, near-coincident or spaced microphone pairs, but they all rely on one optimum configuration at the reproduction side which comprises two loudspeakers positioned in front of the listener with a separation of $\pm 30^{\circ}$.

Binaural audio is a radically different concept for two-channel recording and reproduction that also appeared in the 1930's with the use of a dummy-head for acoustic measurements (Snow, 1934) and for sound reproduction (Fletcher, 1934). The idea of this technique is to record a real acoustic scene directly at the ears of a listener or a simulated version of a human subject, i.e., a dummy-head. In principle, the subsequent reproduction of these signals at the recording position produces an acoustic experience identical or very similar to the original acoustic scene with its associated spatial sound characteristics (for a recent overview of binaural technology, see Hammershøi and Møller, 2005). The binaural technique was initially designed for headphone replay, which is a natural way to deliver the original signals independently to each ear. However, the reproduction of binaural signals over a two-loudspeaker setup is also possible using the principle of crosstalk canceling formulated by Atal and Schroeder

[1]'Stereophony' is employed here as a generic term for all sound transmission methods using two or more channels (Steinke, 1996) and the term 'stereo' is used in this thesis to denote the specific case of two-channel stereophony.

[2]The term 'surround sound' is also employed for this type of recording/reproduction techniques. Steinke (1996) characterized surround sound reproduction as the addition of a diffuse ambience or effects at the back to the more directly localizable frontal sound sources.

[3]In this reference, the concept of binaural technology referred to the pickup, transmission or reproduction of two signals. The modern understanding of the term 'binaural' goes beyond this early interpretation.

(1966). The binaural technique provides good spatial realism over headphones but it suffers from several practical limitations such as a listener dependent performance due to the individual nature of Head-Related Transfer Functions (HRTFs) and a limited compatibility between headphone and loudspeaker replay. Static binaural reproduction[4] poses an additional problem relating to head movements of listeners. These movements produce an associated rotation of the entire audio scene in a headphone replay scenario or a collapse of the binaural effect in a loudspeaker replay scenario. However, dynamic head-tracking solutions have been developed to address this issue in more recent applications of binaural technology as illustrated for example with the binaural room scanning technique reported by Mackensen *et al.* (1999) and Pellegrini (2001).

Multichannel (or surround) sound systems comprise more than two channels and allow therefore the exploitation of a third dimension for the recreation of spatial sound characteristics. The addition of surround channels to the front channels has been historically associated with matrixing techniques such as Dolby Surround and ProLogic designed to emulate cinema sound in consumer applications. Introduced in the 1980's, these two systems work by encoding the center channel and surround channel of the 4-channel Dolby Stereo system into its left and right channels. This approach solved the issue of multichannel storage and delivery while maintaining a downward compatibility with the two-channel reproduction format.

The more recent deployment of discrete multichannel sound formats in the consumer market for audio and audio-visual applications has been favored by the widespread acceptance of the so-called '5.1 surround' system. This loudspeaker configuration is intended for the reproduction of six discrete audio channels comprising five full-bandwidth loudspeakers for the frontal and surround channels and a subwoofer for the low frequency enhancement (LFE) channel. The ITU-R Recommendation BS.775-1 (1994) suggests a loudspeaker positioning of $\pm 0°$, $\pm 30°$ and approximately $\pm 110°$ around the listener for the 5-channel configuration[5]. Note that other multichannel configurations such as the 7.1 and the 10.2 systems have also been proposed based on the idea of adding loudspeakers to the standard 5.1 arrangement as described by e.g. Rumsey (2001).

Multichannel recording/reproduction techniques do not always associate strictly one microphone to a loudspeaker, instead they can use combinations of signals to reconstruct the characteristics of the original sound scene, as found for example in the Ambisonic system developed by Gerzon (1973, 1974). In this approach, signals are recorded with a tetrahedral arrangement of four microphones and are sufficient to recreate a sound field with various loudspeaker configurations. Yet another approach that has found a lot of interest in recent literature is the multichannel sound concept known as wave field synthesis (WFS). This technique developed at the University of Delft (Berkhout *et al.*, 1993; Boone *et al.*, 1995) is based on the principle of reconstructing the original sound field around a listener using an array of loudspeakers. In

[4]The term 'static' is employed here to indicate the absence of 'dynamic' cues occurring in natural listening conditions through head movements.

[5]The term '3/2 stereophony' is often used in audio standards for this configuration.

practice, a WFS system can produce very convincing sound source imaging over a wide listening area but it requires a large number of loudspeakers.

3.2.2 The source-medium-receiver model

The analogy of a 'source-medium-receiver' model is exploited now to illustrate different scenarios of spatial sound reproduction encountered in practice. The model considered here is purely acoustical[6] and is defined by the three following components:

- **Source**: An acoustic object at a given location, for example the sound emitted by a person, a musical instrument or any vibratory object, or a group of such objects at different locations. Acoustic objects can be natural or synthetic and their acoustic properties usually give them specific spatial sound characteristics, e.g. the source directivity of a human talker in free-field.

- **Medium**: The acoustic environment in which the sound source propagates. The acoustic waves emitted by the source can be reflected, diffracted and absorbed to an extent that depends on the environment and this results in very different spatial sound characteristics. Examples of acoustic spaces include free spaces such as an anechoic chamber and enclosed spaces such as a listening room, a concert hall or a reverberant chamber.

- **Receiver**: The physical point(s) or object at which the acoustic event resulting from the source-medium interaction is captured. This can be a single omnidirectional pressure microphone but a more accurate representation of the spatial properties of the acoustic event can be obtained with directional microphones and microphone arrays (see e.g. Peltonen *et al.*, 2001). The receiver can also be a real listener or any physical object affecting the acoustic waves by its presence, e.g. a artificial head as often employed in the binaural recording technique. In most cases, the spatial sound characteristics associated with this receiver are largely affected by its location and orientation.

The physical sound field resulting from this 3-step process can be either perceived directly by the listener acting as a receiver or be recorded and stored for future reproduction. Considering this latter case, it should be noted that the purely acoustic source-medium-receiver model presented above applies similarly to a sound reproduction situation, e.g. a multichannel speaker setup in a listening room. The elements associated to the model include in this case a set of audio sources (loudspeakers), an acoustic environment (listening space) and a listener. This basic model can therefore be combined to illustrate the different scenarios of spatial sound recording and reproduction, as described next.

[6]Begault (1994) employed a similar approach to describe spatial hearing and included in his source-medium-receiver model the hearing system from the ear to the final stages of perception by the brain. The latter perceptual component is not considered in the present model.

3.2.3 Sound (re-)production scenarios

Primary acoustic event

Figure 3.1(a) illustrates a straightforward application of the source-medium-receiver model in which a musical performance is attended by a listener at a concert hall venue. The source in this case can involve a large number of sound events, for example several soloists and an orchestra, which creates a complex set of spatial sound characteristics. In addition, the medium has a major influence on this sound events as illustrated by the research published in the domain of concert hall acoustics (Kuttruff, 1976; Beranek, 1996). Finally, the position of the receiver, for example a listener, has also an impact on the resulting sound field (Hawkes and Douglas, 1971).

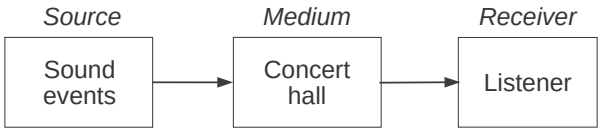
Loudspeaker reproduction of a recorded acoustic event

The reproduction of a recorded acoustic event over a loudspeaker setup introduces a secondary acoustic event in the sense that two paths of the source-medium-receiver model are needed as illustrated in Figure 3.1(b). The first path corresponds to the recording side of the chain and is similar to the scenario described above except that the receiver is now the microphone system employed for the recording of this acoustic event. The second path corresponds to the reproduction side of the chain and includes a loudspeaker setup as a source, a second acoustic space as a medium and a listener as a receiver at the end of the chain. The complexity introduced by this double acoustic path compromises obviously the accuracy of the reproduced spatial sound characteristics and specific recording or reproduction techniques or recording/reproduction combinations might affect the process in specific ways.

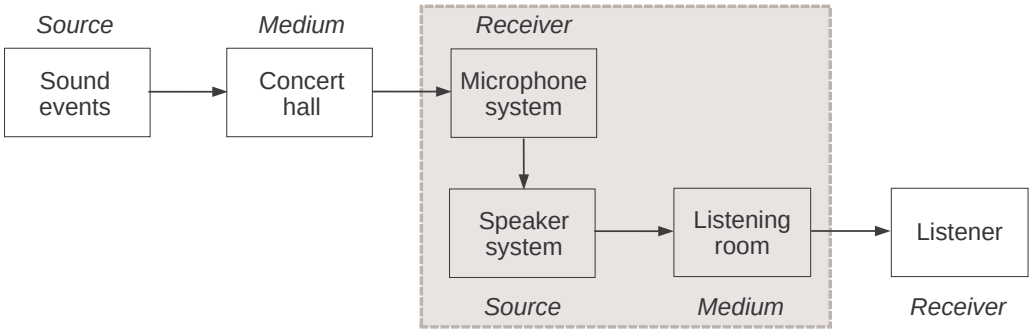
If the goal in this type of scenario is to reproduce as transparently as possible the real acoustic event, the three following blocks in Figure 3.1(b) need to be controlled: the microphone system (receiver – path 1), the loudspeaker system (source – path 2) and the listening space (medium – path 2). The evaluation of recording/reproduction techniques and their optimization towards this goal represent an important research topic of audio engineering. For example, the relevant factors to consider in sound recording and the possibilities and limits of the 5.1 audio format have been reviewed by Theile (2001). The study of the impact of listening spaces on reproduced sound has also been extensively studied (Toole, 1990; Bech, 1994) and reference listening rooms have been defined in audio standards (IEC 60268-13, 1998; EBU 3276, 1998; ITU-R BS.1116-1, 1997; EBU 3276-1, 1999) to control this factor of the chain.

Headphone reproduction of a recorded acoustic event

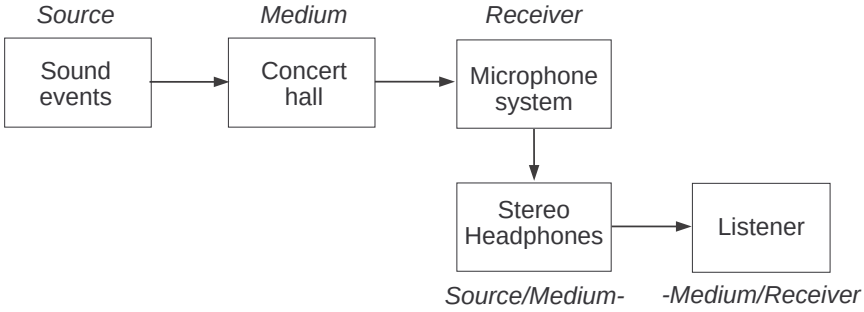
Reproducing an acoustic event over a pair of headphones can be viewed as a variation of the loudspeaker reproduction scenario, as illustrated in Figure 3.1(c). The major difference concerns the reproduction side for which it is not possible to separate the three components of the source-medium-receiver model. The source is physically linked to the receiver and the headphone system placed at the ears of the listener acts as a medium. On the one hand, the reproduction of spatial characteristics is simplified by the absence of spatial interaction with the surrounding space but on the other



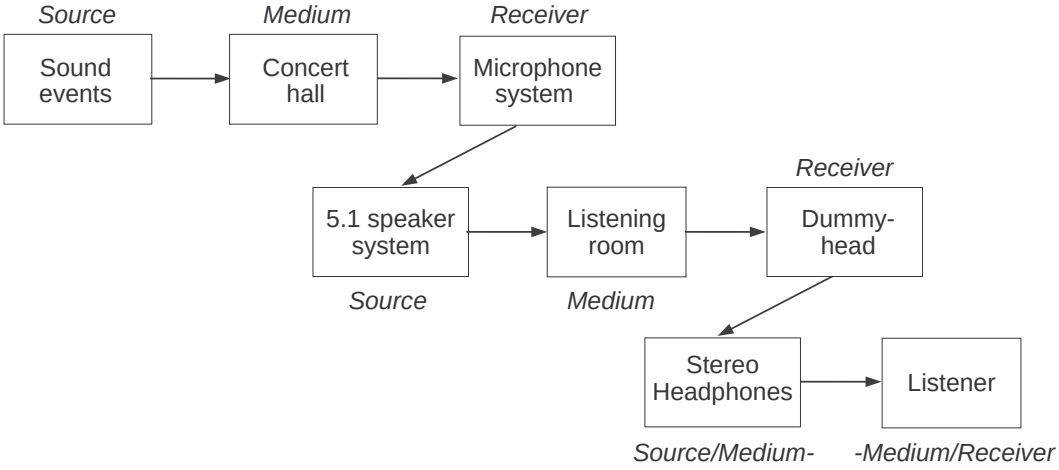
(a) Natural acoustic event



(b) Loudspeaker reproduction of an acoustic event



(c) Headphone reproduction of an acoustic event



(d) Virtual loudspeaker reproduction of an acoustic event over headphones

Figure 3.1: Four examples of spatial sound reproduction scenarios.

hand the source-medium-receiver combination implies very different physical, acoustical and subsequently psycho-acoustical characteristics for this type of reproduction. While headphones are optimal for the reproduction of binaural audio recordings as noted above, they are however in principle not suited for the replay of acoustic events recorded with conventional stereo techniques due to the inappropriate interaural time differences between stereo signals delivered at each ear (Toole, 1984). Bauer (1961) already considered this compatibility issue with stereophonic sound reproduction over headphones in the 1960's and proposed a conversion method based upon the introduction of interaural-time and -level differences between the left and right signals. Related approaches such as the stereo widening network of Kirkeby (2002) have also been developed more recently but the replay of 'unprocessed' stereo content over headphones remains prevalent in mainstream audio applications.

The transparent reproduction of a real acoustic event over headphones is usually compromised by the incorrect acoustic image localization occurring in this special listening scenario. However, research efforts have been made to define the timbral characteristics of headphones that would produce a 'natural' listening experience. The IEC 60268-7 (1996) and ITU-T P.57 (2002) standards define the two following types of reference fields that high-fidelity headphones can replicate: free-field (FF) and diffuse-field (DF). The FF equalization aims to replicate the ear signals produced by a loudspeaker directly in front of the listener. In practice, a procedure of loudness matching to a free-field loudspeaker source can be applied to determine the frequency response correction needed to obtain a FF equalized headphone. The DF equalization uses for reference a sound field in which the angles of sound incidence are equally distributed over the sphere around the listener. The association model proposed by Theile (1986) supports the idea that a DF equalization is better suited than a FF equalization due to the fact that spectral distortions contained in the free-field curve are perceived as an unnatural coloring of the acoustic image when the correct localization does not occur, which is usually the case in headphone listening. It should be noted finally that the topic of headphone equalization is also relevant to binaural sound reproduction and has been investigated for example by Larcher *et al.* (1998).

Headphone-based virtual loudspeaker reproduction of a recorded acoustic event

The generation of a tertiary acoustic event, that is, a scenario requiring three paths of the source-medium-receiver model as illustrated in Figure 3.1(d), can also be conceived in some sound rendering applications. For example, the technique of 'virtual 5.1-channel' reproduction over headphones included in the investigation reported later in this thesis belongs to this category. The aim of this type of digital signal processing algorithms is to reproduce faithfully the spatial characteristics of audio material available in the 5.1 format with a pair of stereo headphones[7].

An approach commonly adopted consists in recreating a 5-channel loudspeaker

[7]Note that in addition to headphone reproduction, the concept of virtual 5.1 can be applied to stereo loudspeaker setups (see Zacharov and Huopaniemi (1999) for a subjective evaluation of such systems) and to wave field synthesis array systems (Boone *et al.*, 1999; Theile *et al.*, 2003).

setup virtually over headphones using the binaural processing technique. In practice, a set of binaural room impulse responses (BRIR) has to be measured with a dummy-head recording system for each loudspeaker of the 5-channel setup in a reference listening room. These digital filters can be convolved at a later stage with any 5-channel material for the synthesis of virtual speakers and the resulting binaural signals are combined for the subsequent replay over headphones. To decompose this full recording/reproduction chain, it appears that three paths of the source-medium-receiver model are needed, as illustrated in Figure 3.1(d). The first path concerns the recording side of the chain and follows the idea of the previous scenarios. The second path corresponds to the virtual reproduction of the multichannel loudspeaker system and is similar to the loudspeaker scenario (Figure 3.1(b)) except that an artificial head is employed as a receiver in this case. Finally, the third path concerns the headphone reproduction side which can also be described with the source-medium-receiver model as discussed above (Figure 3.1(c)).

The reproduction transparency in this scenario is compromised by the added level of complexity introduced by the combined requirements set on the loudspeaker and headphone paths. However, the practical limitations inherent to static headphone replay can be solved by incorporating in the second path a dynamical adaptation of the virtual loudspeaker rendering accounting for listeners head movements as found in the binaural room scanning technique (Mackensen *et al.*, 1999) using a head-tracking system. One of the advantages of introducing a virtual loudspeaker path in the chain is to decouple the recording path from the headphone replay path, which adds flexibility to this multi-stage audio rendering technique.

3.3 Spatial sound perception

3.3.1 Comparison of the sound reproduction scenarios in perceptual terms

The scenarios presented in the previous section give an illustration of the variety of acoustic paths that can be taken to reconstruct the original acoustic event. The listener will ultimately form an auditory sensation from the physical sound field (s)he experiences and it is therefore important to appreciate the level of complexity of these different scenarios in perceptual terms.

While the first scenario involving a primary acoustic event (Figure 3.1(a)) is characterized perceptually by a complete dynamic auditory experience, the three sound reproduction scenarios are only able to recreate this acoustic event to some extent with their specific perceptual limitations. The loudspeaker scenario (Figure 3.1(b)) can simulate some spatial aspects such as the distance and depth of sound events but the stability of the perceived sound image appears to be sensitive to listener position and is restricted in terms of dynamic cues especially in the case of a stereophonic loudspeaker system. The (static) headphone scenario (Figure 3.1(c)) can provide a convincing three-dimensional sound experience when appropriate binaural recordings are employed but in the more common case of stereo sound reproduction, frontal

sound imaging is seriously compromised due to limitations in localization specific to headphone sound perception to be discussed in the next section. The last scenario described above (Figure 3.1(d)) addresses the sound localization issues of the stereo headphone replay by incorporating a binaural component in the second path of the acoustic chain. However, the lack of dynamic information still prevents an accurate frontal localization of the sound image. Thus, only the more complex version of the virtual loudspeaker reproduction scenario using dynamic head-tracking can solve most of the problems specific to headphone sound reproduction, but it should be noted that the performance of this audio rendering technique is then equivalent to an optimal loudspeaker reproduction scenario, which means that it is still dependent on the perceptual limitations introduced by the initial recording step (path 1 of this scenario).

This brief analysis highlighted some essential perceptual differences between the three selected scenarios and illustrated the difficulty to achieve a faithful perceptual reconstruction of the original acoustic event with the techniques commonly applied for spatial sound recording/reproduction. A detailed perceptual evaluation of these techniques appears therefore as a needed component if a complete understanding of this domain is desired.

3.3.2 Perceptual evaluation of spatial sound

The sound recording/reproduction scenarios discussed earlier highlighted the challenge of a transparent reproduction both in acoustical and perceptual terms. Furthermore, in many audio applications the original sound event is impossible to reproduce acoustically or cannot be perceived as identical. Pellegrini (2001) studied this issue in the context of auditory virtual environments and recognized the challenge of an authentic reproduction, that is, the creation of an indistinguishable copy of a real environment. He proposed the less restrictive goal of 'plausibility' defined in terms of suitability for a given application and measured in the perceptual domain. This approach sounds reasonable considering the fact that the reference itself tends to lose its meaning in many sound reproduction scenarios. For example, the imitation of a concert hall acoustic experience in a small room or the simulation of a cinema sound experience over headphones are somewhat bound to produce an artificial perceived experience. Rumsey (2002) also argued that "*reproduced sound and synthetic auditory scene creation can give rise to subjective attributes either not encountered or not considered relevant in natural acoustics*". In all these cases, a perceptual evaluation aiming at measuring the distance between a reference and its reproduced copy is challenging and the need to identify the perceptual characteristics of the audio systems under investigation becomes therefore crucial for the optimization of these systems.

The importance of the descriptive approach for spatial sound evaluation has been acknowledged in the literature and various systems have been proposed to characterize either the whole perceptual domain of spatial sound or more restricted sub-domains or sound applications. Letowski (1989) proposed a 'multilevel auditory assessment language' (MURAL) to characterize sound which comprises the two main categories of 'timbre' and 'spaciousness' and several sub-categories. Rumsey (2002) presented a

scene-based paradigm for the evaluation of spatial audio reproduction characteristics. The standard ITU-R BS.1284-1 (2003) recommends also the use of a set of attributes specifying the quality of two-channel stereophonic and multichannel sound in detail.

In a broad sense, the perceptual domain of spatial sound can be conceived as a layered structure of descriptive components from which a subset of attributes might arise in a given spatial sound scenario. Research has been carried out in many of these perceptual sub-domains but more exploration is needed to cover the whole domain and possibly elaborate a 'sound wheel', that is, a comprehensive terminology describing in a complete manner from a perceptual viewpoint all the sound reproduction scenarios.

3.3.3 Selected literature on spatial sound perception

A literature review of spatial sound studies employing a descriptive analysis approach is proposed in this section. The non-exhaustive selection of sixteen publications presented in Table 3.1 (at the end of this chapter) aims at covering a representative set of research works reported over the last 50 years to illustrate the variety of stimulus types and perceptual evaluation methods employed in the field of spatial sound. The focus of this presentation is placed exclusively on the descriptive measurement side of the reviewed studies, leaving out the hedonic measurement or the link to instrumental measures described in some of these publications. Table 3.1 is organized in five columns summarizing the relevant information from this perspective, that is, the reference of the publication in the first column, the stimulus characteristics in the second column, the perceptual evaluation method in the third column and the attributes elicited either in the form of descriptive terms (in the fourth column) or perceptual dimensions (in the fifth column) depending on the experimental methodology.

The first column of this table offers a chronological view of the selected research works. Note that a larger number of recent studies has been included in this review to reflect an active ongoing research in the field of perceptual audio evaluation both in terms of spatial sound applications and applied test methodologies.

The second column specifies the type of audio material and the associated acoustic chain considered in the different studies. Based on this information, it appears that the selected publications can be loosely split in three groups. The first group involves field studies relating to the domain of concert hall acoustics, that is, experiments made directly in large acoustic spaces with live music performances such as the works of Beranek (1962), Hawkes and Douglas (1971), Barron (1988) and Kahle (1995). Laboratory studies aiming to reproduce concert hall acoustics form a second group and include the work of Wilkens (1975) using binaural recording and headphone reproduction as well as the work of Lavandier (1989) simulating room acoustics with headphone or loudspeaker reproduction. The remaining publications in Table 3.1 can be assigned to a third group concerned more specifically with sound recording/reproduction techniques. Studies in this group can either focus on one side or both sides of the recording/reproducting chain. The former type includes the work of Nakayama *et al.* (1971) on sound reproduction with a different number of loudspeakers, the work of Gabrielsson and Sjögren (1979) with various models of loudspeakers,

headphones and hearing aids, the works of Toole (1985) and Lorho (2007) with different mono and stereo loudspeakers, the works of Guastavino and Katz (2004) and Choisel and Wickelmaier (2005) with different loudspeaker configurations, the work of Hegarty *et al.* (2007) with different 2-to-5.1 channel upmixing algorithms and the work of Kim and Martens (2007) with different recording techniques. The latter type is illustrated by the studies of Berg and Rumsey (1999a,b, 2000) and Zacharov and Koivuniemi (2001a), both of which included a large number of technical factors from the two sides of the recording/reproducting chain with the aim of stimulating a wide range of perceptual aspects for the verbal elicitation process.

The third column of Table 3.1 summarizes the experimental methods applied in the reviewed studies. It appears that they can be loosely attributed to two groups of the 'descriptive sensory analysis' category defined in the classification presented in Figure 2.5 (Chapter 2), that is, the group of verbal elicitation methods and the group of indirect elicitation methods.

Fourteen studies would fit best the verbal elicitation group, although only seven experiments can be formally classified in the two sub-categories defined in Figure 2.5, namely, the CV approach for two studies and the IV approach for five studies. The term 'provided attributes' is employed in the table for the seven other studies of this group because the principle of elicitation process defined in the descriptive sensory analysis category does not apply in these cases. By imposing attributes, these studies circumvent the process of direct elicitation by the assessors but the risks associated with this approach need to be carefully assessed. Firstly, assessors might not interpret the attributes and use the descriptive scales in an consensual way. The validity of the resulting attribute data should therefore be assessed carefully, for example with some of the multivariate data analysis techniques described in Chapter 7 of this thesis. Secondly, a vocabulary selected by the experimenter might not be representative or even suitable for the stimulus set under study, i.e., it could either miss some perceptual aspects or include too many redundant attributes.

Three studies listed in this table belong to the indirect elicitation group presented in Section 2.3 (Chapter 2), that is, the works of Nakayama *et al.* (1971) and Lavandier (1989) based on the MDS technique and the work of Choisel and Wickelmaier (2005) using the PSA method.

The perceptual description resulting from the works described in Table 3.1 can take different forms depending on the applied perceptual evaluation method. The forth column of Table 3.1 contains the list of descriptors employed for most of the experiments from the verbal elicitation group. The descriptive terms of this review are valuable as such because they form a broad lexicon that can be exploited in new perceptual studies. In most studies however, the set of descriptors employed is large and does not give a concise view of the perceptual space under study. This is the reason why in some of these works a smaller set of perceptual dimensions was derived from the gathered attribute rating data through various multivariate data analysis techniques reported in the column 3 of the table. The fifth column presents the interpreted perceptual dimensions for seven of the verbal descriptive studies as well as for the experiments using an indirect elicitation method.

A thorough comparison of the perceptual studies presented in this review is not readily applicable due to the distinct nature of the stimuli and the difficulty of relating formally the applied vocabulary or interpreted perceptual dimensions of the different experiments. However, by exploring on a simple semantic basis the large list of attributes presented in the fourth and fifth columns of Table 3.1, it appears firstly that spatial sound perception is complex and multidimensional in nature and secondly that a broad separation in several perceptual groups can be applied as will be briefly presented next.

Spatial attributes form the largest group and can be loosely split in two sub-groups relating respectively to the perception of sound source(s) and space. Source-related attributes found in this review include *distance*, *width*, *stereo impression*, *elevation*, *localization*, *source width*, *source depth*, *strength and extension of source*, *sound stage width* and *sense of movement* while space-related attributes include *reverberance*, *reverberation*, *room perception*, *envelopment*, *spatial impression*, *spaciousness*, *surround*, *sense of space*, *broadness* and *presence*. It can be noted that both sub-groups contain attributes relating to several specific perceptual aspects which might themselves be seen as lower-level perceptual groups. The level of complexity hinted by this informal classification of spatial attributes has also been described by Rumsey (2002) who made a distinction between source(s), ensemble and environment for the definition of perceived width, distance and depth in the 'scene-based' terminology he proposed for reproduced spatial sound perception.

Timbral aspects are also well represented in the Table 3.1 with attributes such as *bass*, *mid-range*, *treble*, *brightness*, *coloration*, *tone color*, *timbral balance*, and *frequency spectrum*. This group describes sound characteristics distinct from the group of spatial attributes but their presence in the list highlights the importance of timbre in the broad domain of spatial sound perception.

Another group not directly related to spatial aspects might also be formed from this list to describe timbral discrimination (by opposition to spatial discrimination) with attributes such as *blend*, *distinctness*, *clearness*, *muddy/clear* and *Pasty*. Additionally, many other perceptual characteristics can be identified from the reviewed works including for example loudness aspects with the terms *loudness*, *strength* and *emphasis*; temporal aspects with the terms *attack*, *precision*, and *dynamic*; or artefact aspects with the terms *distortion*, *noise*, and *disturbing sounds*.

3.4 Headphone sound perception

Following the broad overview of spatial sound reproduction and perception provided in the two previous sections, the specific topic of headphone sound reproduction is addressed in the last part of this chapter. Some background on headphone sound highlighting the specifics of headphone sound perception is provided first. The headphone sound application of interest in this work is then introduced, that is, spatial enhancement techniques for high-quality audio material reproduced over headphones. Finally, a set of listening experiments conducted by the author to assess such systems is briefly presented as an introduction to the related perceptual studies reported in the following chapters.

3.4.1 Overview

The separate replay of audio signals directly at the ears of a listener with a pair of headphones or earphones results in a very unnatural acoustic and psychoacoustic experience. In comparison to loudspeakers however, this type of sound reproduction offers the advantages of removing the interaction with the surrounding acoustic space and allowing either totally independent or identical sound signals to be presented at the ears of the listener[8]. For these reasons, headphones are commonly employed in psychoacoustic experiments in which the sound signal replay has to be controlled very accurately. A large number of spatial hearing studies based on headphone reproduction of identical or non-identical input signals at the two ears have been published (see Blauert (1997) for a review). Headphone studies have also been reported on perceptual aspects commonly encountered with this type of sound reproduction, that is, 'inside-the-head' locatedness (Blauert, 1997) versus externalization (Durlach *et al.*, 1992), lateralization versus localization (Jeffress and Taylor, 1961; Toole and Sayers, 1965; Plenge, 1974) and headphone sound localization issues such as front-back confusion (Carlile, 1996), or elevation and distance perception (Begault, 1992).

Inside-the-head locatedness is commonly encountered in headphone reproduction[9]. Sound sources produced in a natural context around a listener are perceived outside the head but the atypical acoustic path associated with headphone listening results often in a spatial localization of sources inside the listener's head. By opposition, an acoustic image is said to be 'externalized' if the sound source appears to the listener to lie outside the head. This intracranial perception is typical for stereo material and can also occur with binaural material, especially when virtual sound sources are presented in the median plane. When no externalization is experienced, the term 'lateralization' is employed instead of localization to describe the perception of sound sources inside the head but it should be noted that both types can be experienced simultaneously by a listener (Sayers and Cherry, 1957).

The generation of virtual sound sources outside the listener's head requires signal processing methods that simulate the acoustic filtering occurring in natural hearing situations. The binaural technique already introduced in Section 3.2.1 follows this principle and can be applied with HRTF or BRIR filters. The idea of HRTF processing is to reproduce the free-field acoustic cues at the listener's ears. One essential characteristic of HRTFs however relates to its variability across individuals which is due to the idiosyncratic nature of human ears. This feature is largely problematic for static virtual sound reproduction because most of the localization errors listed above might occur when non-individual HRTFs are employed. Many research works have been published on the topic of HRTFs with a focus on several aspects such as the measurement and perceptual validation of individual HRTFs (Wightman and Kistler, 1989a,b; Riederer, 2005) the study of HRTFs as a function of direction (Asano *et al.*, 1990; Duda and Martens, 1998; Algazi *et al.*, 2001) or the study and modeling of dif-

[8]The terms 'monotic', 'diotic' and 'dichotic' refer respectively to a sound signal presented at one ear only, the same sound signal presented at both ears, and different sound signals presented at each ear.

[9]This problem can also occurs in loudspeaker replay under specific conditions as illustrated by Toole (1970).

ferences between individual HRTFs (Kistler and Wightman, 1992; Brown and Duda, 1998; Middlebrooks, 1999). Distance perception can be improved in virtual sound reproduction when acoustic cues relating to the listening space are present. Begault *et al.* (2000) for instance found that the use of BRIRs increases the proportion of externalized distance judgement in comparison to HRTFs.

It should be noted finally that headphones employed for binaural sound reproduction need to be considered carefully as they have their own complex characteristic transfer function which also shows some variability between individuals and positioning at the ear of the listener. This transfer function needs to be compensated accurately when personalized HRTF processing is intended. More generally, several types of headphones exist that have very different frequency responses (Toole, 1984) and this can affect significantly the timbral quality of the reproduced sound.

3.4.2 Headphone sound applications

In addition to the use of headphones for various types of auditory experiments, this sound reproduction technique is employed in a wide range of audio applications including simple monotic speech communication, virtual auditory displays and complex mobile augmented reality applications (Härmä *et al.*, 2004). However, the most important headphone application nowadays is definitely music listening. The introduction of 'personal stereo' in the 1970's[10] demonstrated the concept of listening to music while being mobile. The use of stereo headphones for music listening with personal stereo players became very popular in the 80's and digital technology favored this development with the deployment of the Minidisc, portable CD and MP3 players. Current mobile multimedia devices can handle high-quality stereo and even multichannel audio material and they often employ advanced signal processing techniques to enhance the spatial quality of sound reproduced over headphones.

These different headphone sound applications require specific perceptual evaluation methods to match different objectives and the definition of an appropriate reference acoustic event to imitate is often a challenge. Two types of experiments have been applied in the literature to address this issue of perceptual evaluation of headphone applications, that is, an integrative evaluation approach relying on a measure of overall impression as illustrated below or a descriptive analysis approach as reported for example by Gabrielsson and Sjögren (1979), Silzle (2002) or Marui and Martens (2006).

3.4.3 Perceptual evaluation of spatial enhancement systems for headphones

To complete this chapter, a set of listening tests performed by the author to assess the quality of spatial enhancement systems for headphone reproduction is briefly presented. The approach selected to assess the perceived benefit of these systems in the three experiments is based on a measure of overall impression with a hedonic question. Recalling the classification of measurement methods presented in Chapter 2 (Figure

[10]The Stereobelt patented by Pavel (1978) was the first portable personal stereo audio cassette player and the Sony Walkman released in 1979 was the first commercial personal stereo.

2.2), this evaluation method can be said to belong to the 'affective domain' of sound characterization. Note also that this summary serves also as an introduction to the research work reported later in this thesis since the same systems will be explored from the 'sensory domain' viewpoint with a descriptive analysis approach in Chapters 6 and 9.

Evaluation of headphone spatial enhancement systems for stereo music reproduction

The first experiment reported in Lorho *et al.* (2002) examined the quality of spatial enhancement algorithms aiming to improve the spatial characteristics of stereo music reproduced over headphones. Nine state-of-the-art systems were collected and applied to six different stereo music clips chosen to provide a wide range of music content, style and recording techniques. Twenty four naive subjects compared the algorithms against each other and against the unprocessed stereo material for each musical clip using a preference ranking procedure.

To summarize the results of this listening experiment, a box plot of the preference ranks is given in Figure 3.2(a) for the nine systems and the unprocessed stereo material based on 144 series of ranking, i.e. 24 subjects times 6 music clips. This graph shows that none of the systems were preferred to the unprocessed stereo music on average. The systems #1 and #2 were not significantly different from the unprocessed stereo material (system #0) but significant differences were observed between some of the systems under study.

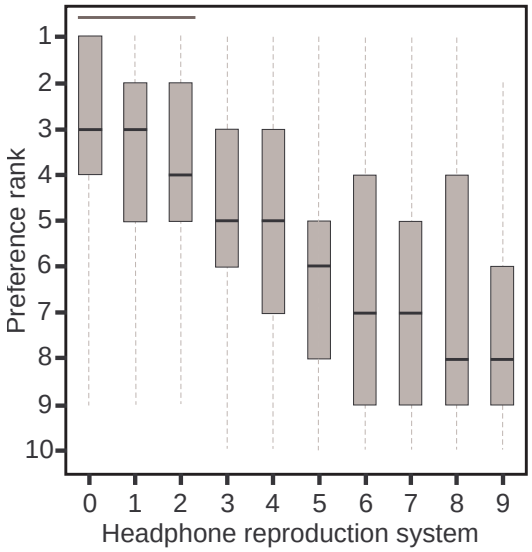
Evaluation of virtual 5.1 systems for headphone reproduction

The second experiment reported in Lorho and Zacharov (2004) examined virtual 5.1 systems for stereo loudspeakers and headphones. These systems aim at reproducing the spatial quality of 5.1 loudspeaker systems and can be applied to different types of high-quality multichannel audio material such as music, movie sound or gaming sound. A separate listening test was run for the loudspeaker and headphone reproduction scenarios as these require different signal processing solutions. Seven virtual 5.1 systems were selected for the headphone experiment and were applied to six different multichannel audio programs including three music tracks, two movie sound tracks and one gaming sound track. Details on the algorithms and associated settings can be found in Lorho and Zacharov (2004). Twenty screened listeners participated to the headphone test for which a preference paired comparison method was chosen using the stereo downmix of the 5-channel material system as a reference.

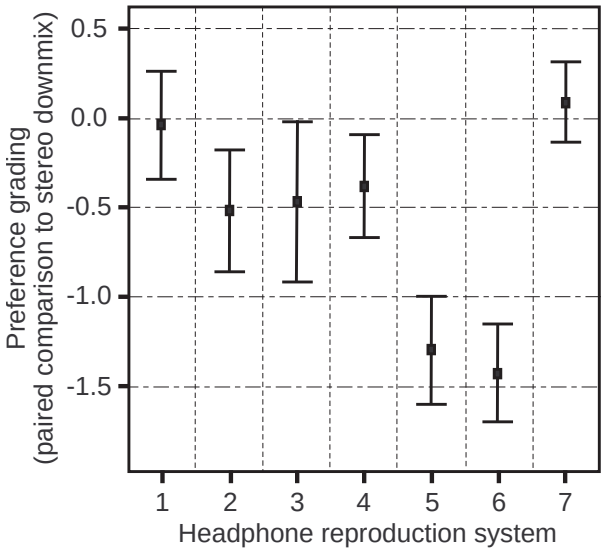
The mean ratings averaged over six program items and twenty listeners is shown in Figure 3.2(a) with 95% confidence intervals for the seven systems. This graph illustrates that none of the systems were preferred to the stereo downmixed material and algorithms #1 and #7 were judged equivalent to the reference on average.

Evaluation of headphone spatial enhancement systems for mono and stereo music reproduction

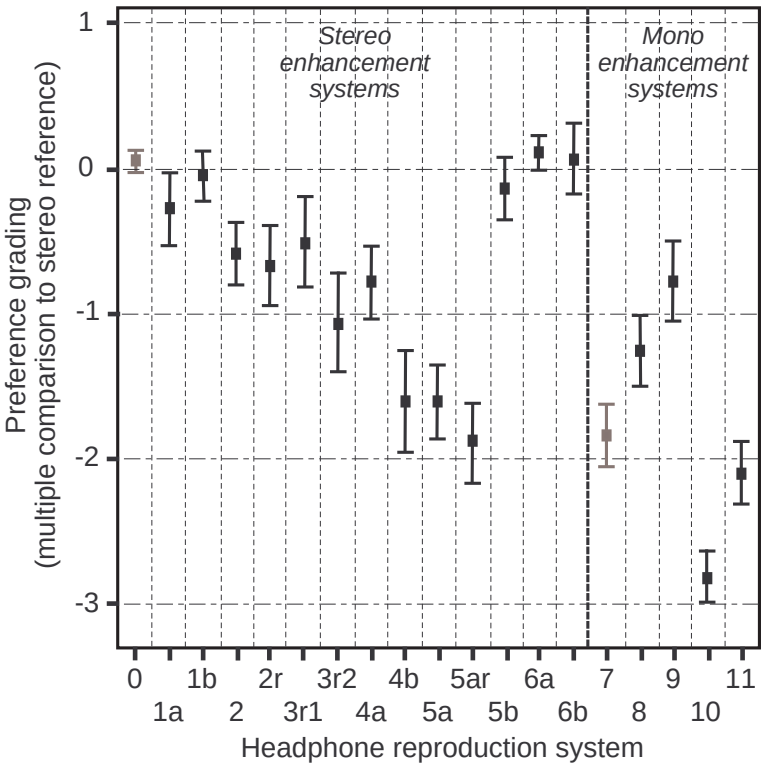
The third experiment reported in Lorho (2005a) examined a larger set of spatial enhancement systems for music reproduction over headphones. Algorithms for stereo



(a) Summary of preference ranks given for 9 headphone spatial enhancement systems and the unprocessed stereo material (System #0). Data represent 144 series (24 subjects × 6 program items).



(b) Summary of preference scores given for 7 headphone virtual 5.1 systems in a paired comparison test using the stereo downmixed material as a hidden reference. Data represent 120 series (20 subjects × 6 program items).



(c) Summary of preference scores given for 19 different systems in a multiple comparison test using the stereo material as reference. Systems include the stereo material as a hidden reference (System #0), 13 headphone stereo enhancement systems, the mono downmixed material (System #7) and 4 mono enhancement systems. Data represent 110 series (22 subjects × 5 program items).

Figure 3.2: Summary of a perceptual study of spatial enhancement systems for headphone reproduction including three subjective preference experiments.

music processing similar to those studied in Lorho *et al.* (2002) were selected but most of them included two different settings. In addition, several mono-to-3D algorithms were considered to widen the perceptual range of spatial differences. This selection gave a total of seventeen systems covering well the styles of spatial widening, room impression and timbral effect that can be obtained with this type post-processing algorithms. Five representative stereo music excerpts were selected for this experiment. The stereo unprocessed material was also mixed down to a monophonic format which was employed in the test and processed with the mono-to-3D algorithms under study. A multiple item rating approach using the stereo unprocessed material as a reference and a 9-point, one decimal place, hedonic scale were employed for this experiment and twenty two screened listeners participated to the listening test.

The summary of the results of this listening test presented in Figure 3.2(c) uses the coding scheme from Lorho (2005a) in which systems denoted from 0 to 6 relate to the stereo-based systems, with 0 being the unprocessed stereo material, and systems coded from 7 to 10 relate to the mono-based systems, with 7 being the unprocessed monophonic material. In addition, the letters a and b refer to distinct algorithms settings and the letter r indicates algorithms including a room effect component (the systems $3r1$ and $3r2$ correspond respectively to a smaller and larger amount of reverberation).

The mean ratings averaged over five music clips and twenty two listeners is shown in Figure 3.2(c) with 95% confidence intervals for the nineteen systems. This graph illustrates that none of the systems were preferred to the stereo material. However, several systems were judged equivalent to the reference on average again, that is, the algorithms #1b, #5b and both versions of the algorithm #6.

Discussion on the result of these studies

This set of listening experiments provided a consistent result that needs careful interpretation. The significant differences observed between most of the spatial enhancement systems in the three tests demonstrates the suitability of the hedonic testing approach for this type of headphone sound application. It can also be noted that the unprocessed monophonic downmixed material was graded lower than the original stereo material by two units (on a negative scale of 4 units) in the third experiment (Figure 3.2(c)). As the perceptual differences between these two sound reproduction techniques relate mainly to spatial characteristics, this result illustrates that listeners were sensitive to this perceptual aspect and agreed on their preference judgement in this case.

However, the fact that none of these systems were found to be superior to the unprocessed stereo material in terms of preference in these experiments was a partly unexpected result that raised questions regarding the methodology applied for this type of evaluation. One point highlighted by the authors in the three publications mentioned above concerns the timbral degradation introduced by some of the spatial enhancement algorithms. This aspect was commented as playing a more important role in overall quality judgment than the spatial characteristics. Considering however the large difference in judgment observed between the mono and stereo material, the importance of spatial aspects cannot be ignored. It is therefore surprising that some spatial enhancement systems applying only minor timbral variations to the original

sound material were not preferred to the unprocessed material on average.

To challenge this surprising result, the author made the hypothesis that an interpretation of results based on data averaged over listeners is inappropriate in this context. This idea can be illustrated by an analogy to a food product example. Considering preference for ice cream flavors, it appears that the vanilla flavor is preferred on average but this does not prevent some people from preferring ice creams with other flavors. Broadly, it can be stated that groups of people preferring different flavors exist in this example and the author believes that preference for spatial enhancement systems might follow the same idea. This would mean that an average preference for the unprocessed audio material conceals sub-group preferences for other algorithms.

As a matter of fact, the result of the third experiment indicates the presence of some differences in preference between listeners as discussed in Lorho (2005a). The ice cream analogy would therefore deserve some consideration in this spatial enhancement system evaluation context. To test the existence of clusters of listeners preferring different algorithms in practice, the evaluation method would need to be adapted to include a much larger number of listeners. As a result, the data obtained from this experiment would allow for a statistically robust identification of such clusters.

This type of large-scale hedonic test strategy follows the idea of the quality evaluation approach presented in Chapter 2. The methodology for quality assessment of spatial enhancement systems presented above could be further aligned with this approach by the addition of a sensory characterization step in the evaluation process. This twofold approach would enable the preference mapping framework presented in Chapter 2 (Section 2.2.4) to be applied. Through this process, a 'target' as defined in Section 2.2.1 rather than a 'reference' as discussed above could be identified and utilized to measure objectively the quality of headphone spatial enhancement systems in the sensory domain.

3.5 Summary

In this chapter, a broad overview of spatial sound reproduction and perception was provided. A literature review on perceptual evaluation of spatial sound illustrated the multidimensional nature of this perceptual domain and the wide range of perceptual aspects it covers. The third part of this chapter examined more specifically the topic of sound reproduction over headphones with a review of the perceptual characteristics associated with headphone sound and a short presentation on the quality evaluation of an audio application referred to as spatial enhancement for headphones.

Table 3.1: A non-exhaustive review of studies on spatial sound perception using a descriptive analysis approach.

Author	**Stimulus characteristics**	**Perceptual evaluation method**	**Descriptors**	**Perceptual dimensions**
Beranek (1962)	- Live music - 54 concert halls	- Semantic description based on inteviews with musicians, conductors and music critics. - Quantitative grading with provided attributes developed by the author	18 qualitative terms reduced to 8 continuous scales: *Intimacy, Liveness, Warmth, Loudness of direct sound, Loudness of reverberant sound, Balance and blend, Diffusion* and *Ensemble*.	
Hawkes and Douglas (1971)	- Live music - 4 concert halls - 4 different positions	- Provided attributes - Quantitative grading - 4 (unspecified) listeners - Factor analysis	14 continuous scales (+ 2 hedonic scales) selected from Beranek (1962) study[1]	5 latent dimensions interpreted as *Reverberance, Balance and Blend, Intimacy, Definition* and *Brilliance.*
Nakayama *et al.* (1971)	- Live music recorded with 8 microphones in a concert hall - Sound reproduced with 1 to 8 loudspeakers in an anechoic room	- MDS (plus preference rating) - 10 naïve[2] listeners	n.a.	3 latent dimensions interpreted as *Depth of image sources, Sensation of fullness* and *Sensation of clearness.*
Wilkens (1975)	- Binaural recording of 3 classical music selections in 6 concert halls - Static binaural replay over headphones	- Provided attributes - Quantitative grading - 40 listeners - Factor analysis	16 continuous scales (+ 3 hedonic scales)[3]	3 latent dimensions interpreted as *Strength and extension of source, Distinctness or clarity* and *Timbre of the total sound.*
Gabrielsson (1979)	- Music, speech and soundscapes - Reproduction on monophonic speakers, stereophonic headphones, and hearing aids	- Provided attributes - Quantitative grading - 20 non-naïve[4] listeners - PCA, INSCAL	55 continuous scales elicited from sound engineers, audiologists and hearing-impaired people	8 latent perceptual dimensions: *Clearness/distinctness, Sharpness/hardness – softness, brightness – darkness, Disturbing sounds, Fullness – thinness, Feeling of space, Nearness* and *Loudness.*

Continued on next page

[1] *Live/dead, Cold/warm, Clear/muddy, Dull/brilliant, Even/boomy, Distant/close, Dry/resonant, Blended/unblended, Balanced/unbalanced, Public/intimate, Reverberant/unreverberant, Unresponsive/responsive, Large/small dynamic range* and *Poor/good definition.*

[2] The term 'naïve' follows the classification presented in Chapter 2 and refers to subjects who do not meet any particular criterion for the assigned task.

[3] *Small/large, Unclear/clear, Soft/hard, brilliant/dull, Rounded/pointed, Vigourous/muted, Blunt/sharp, Diffuse/concentrated, Overbearing/reticent, Light/dark, Muddy/transparent, dry/reverberant, Weak/strong, Emphasized /not emphasized Treble, Emphasized /not emphasized Bass* and *Soft/loud.*

[4] The term 'non-naïve' is employed here to indicate that subjects might be influenced in their judgments by their background in music, acoustics or audio technology.

Table 3.1: continued from previous page.

Author	Stimulus characteristics	Perceptual evaluation method	Descriptors	Perceptual dimensions
Toole (1985)	- 12 stereophonic music excerpts - Reproduction over mono/stereo speakers in a listening room	- Provided attributes - Quantitative grading - 2 × 16 non-naïve listeners	15 continuous scales in two groups: *Spatial quality: Sound image definition, Sound stage continuity, Sound stage width, Impression of distance/depth, Abnormal effects, Reproduction of ambiance/Spaciousness/ reverberation, Perspective* and *Overall spatial rating* *Sound quality: Clarity/definition, Sotfness, Fullness, Brightness, Hiss/noise/distortions, Pleasantness* and *Fidelity.*	
Barron (1988)	- Live music - 11 concert halls	- Provided attributes - Quantitative grading - 27 non-naïve listeners	10 continuous scales: *Clarity, Reverberance, Envelopment, Intimacy, Loudness, Balance of* 1) *Treble re. mid-freq.*, 2) *Bass re. mid-freq*. and 3) *Singers/ soloists re. orchestra, Background noise* and *Overall impression.*	
Lavandier (1989)	- 3 music excerpts and 1 speech sample - Anechoic recordings processed with a virtual acoustic system - Headphone or speaker replay	- A series of MDS tests - 12 non-naïve listeners	n.a.	14 interpreted perceptual factors: *remote–present, weak–strong, muddy–clear, distant–near, dry–reverberant, flat–contrasted, flowing–halted, hard–soft, neutral-intimate, dry–lively, hollow–warm, weak–bright, spatial impression* and *source width.*

Continued on next page

Table 3.1: continued from previous page.

Author	Stimulus characteristics	Perceptual evaluation method	Descriptors	Perceptual dimensions
Kahle (1995)	- Live music - 10 concert halls	- Provided attributes - Quantitative grading - 10 non-naïve listeners - PCA	25 categorical scales (+ 5 hedonic scales)	Subset of 8 'important' attributes: *Strength, Reverberance, Overall balance, Contrast, Low frequency strength, High frequency strength, Pasty* and *Halted.*
Berg and Rumsey (1999a, 1999b and 2000)	- Speech, music and soundscape - Recording using different microphone techniques - Reproduction in a listening room using different loudspeaker configurations	- Repertory Grid Technique - 18 non-naïve listeners - VPA[1] and Cluster analysis	342 individual 'constructs'	12 descriptive construct groups identified: *'Localisation, left - right and front - back', Depth/distance, Envelopment, Width, Room perception, Externalisation, Phase, Source width, Source depth, Detection of background noise* and *Frequency spectrum.*
Zacharov and Koivuniemi (2001)	- Several types of sound material - Recording with different microphone techniques in different room acoustics - Reproduction in a listening room with 8 loudspeaker systems.	- Consensus verbal elicitation - Quantitative grading - 12 selected[2] listeners	12 continuous scales. *Sense of direction, Sense of depth, Sense of space, Sense of movement, Penetration, Distance, Broadness, Richness Naturalness, Hardness, Emphasis* and *Tone colour.*	
Guastavino and Katz (2004)	- Ambisonic recording of urban soundscapes[3] - 2D and 3D speaker reproduction in listening room	- Individual elicitation by free verbalization (qualitative) - 27 non-naïve listeners - Grouping by semantic analysis	n.a.	6 perceptual groups identified: *Readability, Presence, Distance, Localization, Coloration* and *Stability.*

Continued on next page

[1] Verbal Protocol Analysis (Samoylenko *et al.*, 1996) is a technique for classification of descriptors in groups based on their nature, e.g. descriptive versus attitudinal in the study of Berg and Rumsey.
[2] The term 'selected' follows the classification presented in Chapter 2 and refers to subjects chosen for their ability to perform a sensory test.
[3] This study refers to 'Experiment 1' in the article of Guistavino and Katz (2004).

Table 3.1: continued from previous page.

Author	Stimulus characteristics	Perceptual evaluation method	Descriptors	Perceptual dimensions
Choisel and Wickelmaier (2005)	- 5-channel music material - Reproduction in a listening room with different loudspeaker configurations	- Repertory Grid Technique (RGT) and Perceptual Structure Analysis (PSA) - 40 selected listeners - Cluster analysis (for RGT)	- A large number of individual descriptors for RGT - n.a. for PSA	'8 perceptual categories occurring most often' after Choisel and Wickelmaier (2007): *Width, Brightness, Spaciousness, Elevation, Distance, Envelopment, Naturalness* and *Clarity*.
Kim and Martens (2007)	- Four solo piano pieces played in one concert hall - Recordings with four surround microphone arrays - 5-channel sound reproduction in a listening room	- Individual verbal elicitation - 8 non-naïve listeners - Attribute subset selected by the experimenters	A large number of individual descriptors	5 descriptors selected 'to represent the most salient differences' between the stimuli: *Width, Distance, Focus, 'Bass Tightness'* and *Sharpness*.
Lorho (2007)	- Stereo music material - monophonic, stereophonic and 3D audio processed reproduction over mobile multimedia loudspeakers	- Individual verbal elicitation - Quantitative grading - 16 naïve listeners - Qualitative grouping based on attribute definitions - Quantitative analysis by PCA	111 individual descriptors	5 perceptual groups interpreted as *loudness*, *spatial aspects* (including *space* and *sound localization*), *timbral aspects*, *sound disturbance* and *sound articulation*.
Hegarty, Choisel and Bech (2007)	- Stereo music material - 2-to-5.1 channel upmixing algorithms - Two reproduction techniques: 5.1 in a car and car auralization in a listening room	- Consensus verbal elicitation performed with the stimuli of the car environment - Quantitative grading - 7 selected listeners	15 continuous scales: *Attack, Surround, Stereo impression, Fullness, Reverberation, Continuous noise, Details, Bass, Midrange, Treble, Loudness, Precision, Distortion, Dynamics* and *Distance*.	

Chapter 4

Overview of consensus vocabulary methods

4.1 Introduction

The group of consensus vocabulary (CV) methods introduced in the classification of sensory analysis techniques presented in Chapter 2 refers to the descriptive sensory analysis techniques based on verbal elicitation and employing a group of assessors to develop a common vocabulary. The CV approach represents one of the most sophisticated tools in sensory science both in terms of methodological procedure and sensory panel involvement. Running this type of experiment requires time, commitment and careful planning in practice but when implemented successfully, it offers a powerful means to explore the sensory space of the stimulus set under study.

As one of the aims of this thesis work was to apply this descriptive analysis approach to develop a consensus vocabulary for headphone sound reproduction, an overview of the CV methodology is proposed in this chapter. Section 2 includes some historical background, a description of the main techniques in this category and a list of features common to these different procedures. The meaning and implications of the essential concept of 'consensus' specific to this methodology are also discussed and the desired characteristics of sensory descriptors are reviewed. In Section 3, issues relating to the validity of a consensus vocabulary are addressed and techniques to evaluate the performance of this type of sensory panel are presented.

4.2 Historical background of consensus vocabulary methods

Describing the sensory characteristics of complex stimuli is a difficult task that has been historically performed by one or several individuals considered as experts in their field, e.g. in the brewing, wine or perfume industry. Through training and experience, these experts developed a terminology to assess products and were able to contribute to some extent to product quality improvement and assurance. However, this expert approach had some major drawbacks such as the reliance on only one or a limited

number of experts, the potential lack of objectivity and robustness in the product description and the practical ineffectiveness in a larger industrial context. All these limitations motivated the development of modern sensory analysis methods which can be characterized by a set of principles contrasting with the traditional expert approach, that is, 1) employing a panel of trained assessors rather than a single expert, 2) developing an objective description of products based exclusively on sensory properties, and 3) performing quantitative evaluations with reliable scales. Flavor Profile (Cairncross and Sjöström, 1950) was the first formal descriptive analysis method to follow this approach and was developed during the late 1940's by researchers working in the field of flavor research at Arthur D. Little laboratory, Cambridge, Massachusetts. The introduction of this technique demonstrated that a small group of people can be trained to describe accurately their perception and this motivated the development of several other methods in the following decades, i.e., Texture Profile (Brandt *et al.*, 1963) in the 1960's, Quantitative Descriptive Analysis (Stone *et al.*, 1974) in the early 1970's, Sensory Spectrum in the late 1970's (Civille, 1979) and more recently Quantitative Flavour Profiling (Stampanoni, 1993). In addition to this list, the approach referred to as 'generic (consensus) descriptive analysis' by Einstein (1991) or Murray *et al.* (2001) can be mentioned. This hybrid method combines different elements of the classical techniques listed above and has found a lot of applications to various product categories in the last decades (see Lawless and Heyman (1998) for a review). The retrospective view on these different consensus descriptive analysis methods presented next will highlight the features common to these techniques and also the evolutions or trends in sensory practices over the years.

The Flavor Profile (Cairncross and Sjöström, 1950; Sjöström, 1954; Caul, 1957) employs small group of assessors (4 to 6 individuals) working essentially as a consensual group. The procedure includes an initial vocabulary development and an intensive training phase of few weeks during which the panel is exposed to a wide range of products to ensure that assessors become sensitive enough to stimulus differences. In practice, the panel has to agree on the intensity of the product for each sensory descriptor and has to describe the order of appearance of the different sensations. In the original Flavor Profile, the consensual scaling is made with a 5-point category scale using symbols with the following word anchors: *not present*, *threshold*, *slight*, *moderate* and *strong*. A score of 5, i.e., *strong*, on this scale represents the highest intensity of an attribute in the whole product category, which means that the sensory measure is intended to be absolute with this sensory analysis technique.

The Texture Profile (Brandt *et al.*, 1963) was developed at the General Foods Technical Center in New-York to describe exclusively the textural properties of food products. The principle of this technique derives from the Flavor Profile but it employs a larger panel (a minimum of 10 carefully screened individuals). The main specificity of this technique relates to the pre-defined descriptive terminology it utilizes. This vocabulary is based on a detailed classification developed by Szczesniak (1963) and includes sensory definitions derived from physical and rheological properties (Szczesniak *et al.*, 1963). Additionally, the intensity scales adopted for the attribute rating are associated with a very detailed set of words and illustrative products distributed on

a 13-point scale. The development of a panel for the Texture Profile method requires therefore a relatively long training period (over 6 months) but it allows in principle an absolute sensory measurement of any food product.

The Quantitative Descriptive Analysis (QDA), developed at the Stanford Research Institute in California, introduced a rupture from the previous descriptive analysis practices by proposing several new ideas (see Stone and Sidel (1993) for a review). Firstly, the panel of screened assessors (usually of a group of 10 to 12 individuals) is not subjected to any external influence, for example from the panel leader, in order to avoid potential bias in a technical direction during the vocabulary development. Secondly, the descriptive terminology aims at covering all the product properties which differs from the Texture and Flavor Profile methods. Thirdly, the sensory evaluation is not intended to be absolute, meaning that the attribute intensity measurement is only relative to the set of products under study, and the use of references is also less systematic. This approach reduces considerably the effort needed for panel calibration, which requires typically 10 to 20 hours of training time in QDA. And fourthly, the attribute scaling is made individually by the assessors using a quantitative unstructured line scale and subsequent data analysis is employed to control the internal validity of the sensory panel.

The Spectrum descriptive analysis method (see Meilgaard *et al.* (1991) for a review) is mainly based on the Flavor Profile and Texture Profile techniques but it also shares some ideas with the QDA approach. The panel consists of 10 to 15 screened assessors who develop technical expertise in the product category through a comprehensive training procedure. A descriptive terminology covering all the product properties is built using a set of absolute category scales which are calibrated to have equi-intensity across scales (e.g. 5 on a sweetness scale is of equal intensity to 5 on a salty scale). Similarly to the Texture and Flavor Profile methods, the scales are based on an systematic use of reference points with corresponding food reference samples but the attribute scaling is made individually by the assessors following the QDA approach.

Quantitative Flavor Profiling (QFP) is the most recent of the methods presented in this review (see Stampanoni (1994) for an application). Developed by Givaudan-Roure in Switzerland to characterize flavor perception, this technique can be considered as a combination of the Flavor Profile and QDA approaches. The procedure proposed in Quantitative Flavor Profiling is characterized by two separate phases. In the first step, a technical descriptive terminology with associated absolute references is developed by a relatively small panel (6 to 8 individuals) of experts in the field. In the second step, a group of 10 to 15 screened sensory assessors is trained to use reliably this standardized flavor vocabulary and the evaluation of the products under study is made similarly to the QDA approach. This two-sided procedure ensures the development of a universal and unambiguous terminology (in step 1) while allowing for a separate product evaluation by trained and unbiased sensory assessors (in step 2). The development of a formal set of physical standard references plays an crucial role in QFP as it ensures both an efficient training of the assessors and an absolute sensory measurement.

Several interesting elements can be highlighted from this review. The principles shared by these consensus vocabulary development methods, which relate to the nature of the panel, the development of a formal descriptive terminology and the use of attribute intensity have already been highlighted above. In addition, it appears that the idea of running an individual sensory evaluation of the products in isolated booths, which was introduced by the QDA method, has found favor in later descriptive analysis methodologies. As a matter of fact, this approach combined with the use of quantitative line scales offers an efficient means to control statistically the internal validity of the sensory panel. Interestingly, the use of a consumer-oriented (or not technically-biased) vocabulary with relative attribute scales seems to be specific to the QDA methodology as all the other techniques considered in this review rely on an expert-oriented terminology with absolute attribute scales. The relative attribute scaling approach used in QDA requires a less intensive panel training process, which can be seen as an advantage in practice, but it also limits possibilities to compare results across products, as discussed for example by Munoz and Civille (1998).

The differences observed between these techniques are important to consider when selecting a methodology for a descriptive sensory analysis program because this has a large impact on both the implementation and the outcome of the project. For example, a one-off sensory evaluation project and a long-term sensory quality control program would require very different vocabulary development approaches in practice. The common application of generic descriptive analysis, which offers the flexibility of selecting and combining elements of the classical methodologies, illustrates also the multiplicity of approaches needed to fit specific project goals.

4.3 Development of a consensus vocabulary

The complexity of the CV approach is apparent from the advanced procedures employed in the different methods presented above. The aim in all these techniques is to develop a reliable sensory measurement system, which comprises a descriptive terminology and a panel of assessors trained to use these terms in a consensual way. Several important aspects of consensus descriptive analysis are discussed next. This includes first a description of the steps commonly found in the development of a consensus vocabulary and its associated sensory panel, then a discussion on the concept of consensus in a vocabulary development, and finally a review of the expected characteristics of a consensus vocabulary.

4.3.1 A three-step process to create a consensus vocabulary

The development of a consensus vocabulary includes few key elements that are common to all the descriptive techniques described earlier. The three following steps can be identified in this process: (1) the panel selection, (2) the consensus vocabulary generation, and (3) the panel training. Note that a large literature is available on these aspects including text books (Meilgaard *et al.*, 1991; Stone and Sidel, 1993; Lawless and Heyman, 1998) and standard recommendations (ISO 8586-1, 1993; ISO 8586-2, 1994; ISO 11035, 1994; Majou *et al.*, 2001; ISO 13299, 2003).

Panel selection is the prerequisite step for the successful implementation of any descriptive analysis program. A screening procedure is necessary to identify subjects who show the required characteristics for a sensory evaluation task. The factors to consider during the screening process include various elements such as the level of sensory acuity of the individual in the modalities of interest, his or her motivation, availability and personality. The verbalization skills of the person and his or her attitude in a team are also important for the vocabulary development process and the group discussions. The sensory panel is chosen from a large pool of subjects and consists usually of 6 to 20 selected assessors[1].

Consensus vocabulary generation aims essentially at developing a common vocabulary to describe the sensory characteristics of the set of products under study. This process includes several phases during which the assessors will explore the stimulus set or product category under study. Most of the time, the vocabulary generation starts with an individual task involving a classification of stimuli into categories based on associations and differentiations and an identification (naming) of the stimulus characteristics. In the next step, the whole group of assessors has to establish a single list of sensory descriptors covering the product space of interest. This part of the process is essential and specific to the CV approach. Under the supervision of a panel leader, the assessors have to find an agreement on the attributes and intensity scales to choose and they also create accompanying definitions and physical references in most cases. This group process is iterative and may require 5 to 10 sessions in practice.

Panel training is the last step of the consensus vocabulary development before the formal sensory evaluation can be performed. This process is intended to ensure a consistent and consensual use of the developed attributes scales by the assessors. Training has a significant impact on the quality of sensory evaluations made by a panel, as discussed for example by Labbe *et al.* (2004), but it is also known to require time and effort. It should be noted that the level of separation between the consensus vocabulary generation and the panel training depends on the methodology. For example, the consensus vocabulary development of the QDA approach can be viewed as a continuous process because the same group of assessors participate to the different steps and the vocabulary can still be refined during the panel training phase. However, this overlap does not happen when the sensory assessors employed for the product evaluation do not develop the vocabulary themselves, as for example in the QFP methodology. In such cases, the panel training requires a larger effort including usually a phase of familiarization to the vocabulary and the intensity scales followed by a number of practice sessions during which the assessors use the attributes and get feedback on their performance.

4.3.2 The concept of 'consensus' in a vocabulary development

The three-step process described above provides a double output in the sense that a formal vocabulary has been established and a group of assessors has been trained to use this set of descriptors. Also, the stamp 'consensus' should in principle only be assigned when such a combination has been achieved. It is important to note however

[1]The term 'selected assessor' follows the assessor terminology presented in Table 2.1 (Chapter 2).

that ensuring a good level of consensus within a sensory panel is not a simple task and for this reason the processes involved in a consensus vocabulary development have deserved a lot of attention in the literature. For example, Munoz and Civille (1998) discussed in this context the concept of 'frame of reference' which they defined as the background information that assessors refer to when using attributes to describe perception (qualitative aspect) and when using intensity scales (quantitative aspect). These authors consider the development of a common frame of reference as being the most important part of a descriptive analysis process because it ensures that the panelists have a common understanding of the attributes employed to describe their perceptions. If this alignment process is not included or applied only partially, assessors will continue to use their own frame of reference to evaluate the products and this will usually result in a large within panel variability.

O'Mahony (1991) has also studied this topic which he refers to as the formation and alignment of sensory concepts[2]. This author insists on the idea that the formation of sensory concepts is in fact a complex mechanism including two successive processes: abstraction and generalization. Colors are simple examples of concepts learned from childhood. For instance, the label 'green' does not refer to a single color but to a category formed by the various shades of green. A child is taught to associate certain labels with certain stimuli, e.g. grass and trees for green, and he gradually forms an abstract sensory concept for that particular color. The process of generalization occurs when the child broadens this concept to other stimuli, for example mint or olives. Sensory concepts encountered in descriptive analysis can be very complex. Also, aligning concepts in a panel is usually a difficult and time-consuming process involving iterative adjustments, splitting or merging of sensory concepts. In practice, concept learning is easier for assessors when they are exposed to reference standards, i.e., products illustrating the concepts and their boundaries. The selection of these physical references during the vocabulary development is therefore essential and it plays an important role in most of the consensus techniques described above (see e.g. Murray *et al.* (2001) for a discussion on this topic).

In addition to the alignment of sensory concepts, quantitative aspects relating to attribute intensity are also addressed during the consensus vocabulary development of most descriptive analysis techniques, especially in methodologies using an absolute attribute rating approach. According to Munoz and Civille (1998), sensory evaluations made without a common quantitative frame of reference are limited to relative comparisons of products and do not allow for comparisons with products outside this range of sensory characteristics. To ensure uniformity in scores among the panelists, criteria for attribute scaling have to be established carefully during the vocabulary development. In practice, the intensity boundaries of the attributes can be defined with physical references, for example samples from the product category under study, which helps assessors to adjust to these reference points[3]. The importance of such references in panel training has been discussed by e.g. Rainey (1986). Munoz and Civille (1998) also highlighted the advantages of selecting absolute reference points for

[2]A sensory concept relates to the term 'sensory attribute' defined in Section 2.3 (Chapter 2).

[3]For the auditory modality, one or several sound stimuli can be selected to illustrate a sensory descriptor. Such physical references are referred to as 'audio exemplars' in this thesis.

attribute scale anchoring. They named this method 'universal scaling' by opposition to two alternative approaches called 'product specific scaling' and 'attribute specific scaling' which only allow for a relative scaling of the products under evaluation.

These qualitative and quantitative aspects of attribute usage represent an essential component of the consensus panel training phase. This issue of 'panel calibration' has deserved some attention in the field of sensory science and novel methods have been developed recently to better understand and optimize this process as illustrated for example by the feedback calibration method proposed by Findlay *et al.* (2007).

4.3.3 Desired characteristics of a consensus vocabulary

Based on the wide range of consensus vocabulary profiling applications found in various fields of human perception, it can be safely stated that developing a reliable sensory measurement tool with a group of assessors using a consensus methodology is a feasible task. It should be emphasized however that in order to achieve this goal in practice, it is important to follow the three generic steps described earlier and to consider carefully the specificities of consensus vocabulary development discussed above. Defining clearly the expected outcome of this process is also a relevant matter. Piggott (1991) gives a concise summary of the desired characteristics of a consensus vocabulary as follows: "*the ideal vocabulary consists of terms (...) that accurately and precisely describe all required characteristics of all the samples, they are understood by the assessors who agree with each other about their meanings, they can be easily defined with reliable standards, and they are understood by the recipients of a report*". An exhaustive list of related requirements for elicited attributes has also been presented by Lawless and Heyman (1998). Additionally, certain types of attributes should be avoided such as terms of attitudinal or hedonic nature, e.g. 'authentic', descriptors that are too general, e.g. 'complete', and intensity terms, e.g. 'mild' or 'strong'. In practice however, identifying the nature of the elicited attributes is not always a straightforward process. For instance the character of the attribute 'naturalness' in spatial audio studies requires extra attention. In this respect, definitions are essential to disentangle the exact nature of the elicited attributes.

Wolters and Allchurch (1994) discussed the potential issues with some requirements for elicited attributes. For example, on the aspect of the discriminative nature of selected descriptors, i.e. the fact that attributes should show sample differences, they insist on the difference between two types of attributes in vocabulary development, namely 'descriptors' and 'discriminators', and they consider that whether an attribute discriminates products or not depends on the sample context. In addition, Wolters and Allchurch (1994) highlight the distinction between correlation and redundancy of attributes. The first term relates to a statistical relationship between attribute ratings while the second term relates more to meaning. The fact that sometimes different sensory aspects co-vary has to be assessed carefully during the vocabulary development and appropriate physical references should be created if possible to illustrate the independence of selected attributes in order to help assessors during the training phase, as described by Lawless and Heyman (1998).

The use of attributes based on singular perceptual elements is a related desired criterion for a terminology. Civille and Lawless (1986) indicated that 'primary' descriptors

reduce confusion among assessors while 'integrated' terms can cause problems as each subject might weight the underlying primary terms differently. For instance, Rumsey (1998) discussed this issue in the context of spatial sound perception indicating that the attribute 'Spatial impression' might be composed of more elementary perceptual aspects such as 'envelopment', 'source width', etc. Also, Neher *et al.* (2006) developed a perceptual evaluation technique to address the issue of attribute unidimensionality in the case of spatial auditory stimuli.

4.4 Validity of a consensus vocabulary

The quality of a consensus vocabulary depends ultimately on the ability of the panel of assessors to act as a reliable measuring instrument. Measuring assessor performance is important as it helps ensuring the validity of a consensus vocabulary/panel. This can also be very useful during the phase of sensory concept alignment in a vocabulary development. A number of publications can be found on the topic of panel performance measurement and new approaches continue to be presented in the field of sensory science. An overview of existing techniques and their application to the field of audio was presented in Zacharov and Lorho (2006) and is expanded in this section.

The issue of panel performance measurement relates closely to the general concepts of 'precision' and 'accuracy' which are briefly recalled below. The specific application of these measures in the context of consensus descriptive analysis is also discussed with first a clarification of the terminology, then an illustration of two complementary perspectives to the problem and finally a review of the methods found in the literature for panel performance monitoring.

4.4.1 Measuring precision and accuracy in sensory analysis

The general concepts of precision and accuracy have been defined in the ISO 5497-1 (1987) and ISO 5725-1 (1994) standards as follows.

The precision of a measurement method relates to the variability between repeated measurements and is defined as the closeness of agreement between mutually independent test results obtained under stipulated conditions. Precision is often stratified into the two different concepts of 'repeatability' and 'reproducibility'. Repeatability is defined in ISO 5497-1 (1987) as the closeness of agreement between mutually independent test results obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within short intervals of time. Reproducibility is defined as the closeness of agreement between test results obtained with the same method on identical test material in different laboratories with different operators using different equipment.

The accuracy of a measurement method is defined in e.g. ISO 5725-1 (1994) as the closeness of agreement between a test result and the accepted reference value.

These generic terms have been defined for any measurement method and can therefore be applied to the field of sensory profiling, considering that a sensory panel is a perceptual measurement system from which reasonable objectivity is expected. How-

ever, the mapping of these concepts of precision and accuracy in the context of sensory analysis is not straightforward for two reasons. First, the measurement method can be understood as a given assessor in a panel or as the group of assessors forming a panel. Then, the definition of an 'accepted reference value' for the measure of accuracy is problematic. Even if physical measurements can relate to sensory attributes in some applications, an instrumental equivalent of the perceptual tool developed for sensory analysis is not available and therefore no 'true value' can be used as a reference for the evaluation of the measurement system accuracy (Mangan, 1992). Acceptability or preference judgments have been suggested (Lawless and Klein, 1991) as an external reference for validation purpose but the most common approach found in the literature is to define an internal reference based on related sensory analysis tools, i.e. by comparing the result of a given assessor to the result of the panel, or by considering the result of a given panel in the context of a larger experiment including several sensory panels as described e.g. by McEwan *et al.* (2002).

It can also be noted that an ambiguity exists between the terms 'precision' and 'accuracy' for the measurement of sensory panel performance as can be seen from the different terminologies used in the literature. For example Hollowood (2004) and Schlich *et al.* (2004) use the term 'accuracy' for the measurement of inter-agreement between assessors while Rossi (2001) only relies on the concept of precision by defining 'repeatability' as the ability of an assessor to score the same stimulus consistently and 'reproducibility' as the ability of an assessor to score the stimuli similarly to the other panel members. Mangan (1992) also argues that the concept of accuracy is inadequate for sensory panel evaluation when the panel mean is considered and she prefers the term 'reproducibility'.

4.4.2 Validity criteria for a consensus vocabulary

Based upon the above review of the two generic terms 'precision' and 'accuracy' and the discussion on issues with the application of these concepts to sensory analysis, a definition is now proposed for the terms **repeatability**, **agreement** and **discrimination**. These three criteria are commonly used for panel performance measurement and will be adopted in the context of this thesis.

Repeatability relates to the concept of precision but concerns only the aspect of intra-agreement of the measurement system, i.e., the closeness of results between several measurements made by the same system. In sensory analysis, the term 'system' can relate to a single assessor or to the panel. Repeatability can therefore be defined at both levels but it can only be measured when replicates are included in the experiment.

Agreement relates usually to the aspect of inter-agreement of measurement systems in sensory science but the definition of the term 'system' depends on the context. If the analysis concerns a single panel, the system can be defined as a single assessor but for a study involving several panels (see e.g. McEwan *et al.*, 2002), it is possible to define each panel as a system. In this thesis, agreement will be defined as the closeness of results between measurements made by different assessors of the same sensory panel.

Discrimination relates to the ability of an assessor or a panel to perceive differences between stimuli with one attribute or a set of attributes. The use of this criterion is somehow specific to sensory analysis and it arises from the fact that an 'accepted true value' is usually not available for the measurement of accuracy as specified in e.g. ISO 5725-1 (1994). Measuring discrimination gives therefore a means to assess the utility of the sensory measure.

The three criteria of repeatability, agreement and discrimination form a very good basis for assessing the performance of a consensus sensory panel. It should be noted that only the two former criteria are independent. The discrimination of a single assessor is linked to its repeatability while the discrimination of a panel depends not only on the repeatability and the discrimination of each assessor but also on the level of agreement between the assessors. These criteria can therefore be considered as nested because no discrimination will be achieved if assessors show a poor repeatability, but even when all assessors are able to perceive clear product differences, discrimination between products might still remain poor at a panel level if a large disagreement exists between the assessors.

4.4.3 Univariate versus multivariate assessment of panelist performance

The interpretation of these three criteria from a univariate and a multivariate point of view are now illustrated with a simulated example (adapted from Schlich *et al.*, 2004). The simple sensory profile selected for this purpose is limited to two assessors, two attributes, three products and three replicates per product, as shown in Figure 4.1. In this graph, each triangle represents the three replicates of an assessor for a given product. The three products are shown with a different symbol (cross, circle and square) and the two assessors are presented with a different color, i.e. solid blue for one assessor and dotted red for the other assessor, which will later be referred to as the assessors S and D respectively.

Univariate view: considering this data from a univariate viewpoint, which means looking at each attribute separately by projecting the data on the two original axes, one perspective on the three criteria defined above can be illustrated. It appears in this example that the assessor D has a better repeatability on the attribute #1 than on the attribute #2 (this is visible from the larger spread on the vertical axis in comparison to the horizontal axis) while the assessor S shows a lower repeatability on both attributes. In terms of discrimination, the assessor D is more discriminative on the attribute #1, whereas the assessor S is less discriminative for both attributes because of his lower repeatability. Finally, a relatively good agreement between the two assessors can be seen for the attribute #1 but some disagreement appears on the attribute #2 for the products 'square' and 'circle'.

Multivariate view: another approach to interpret this simple sensory data set is to consider directly the two-dimensional map. The two arrows shown in Figure 4.1 represent the principal directions of variation in the data and are referred to as the

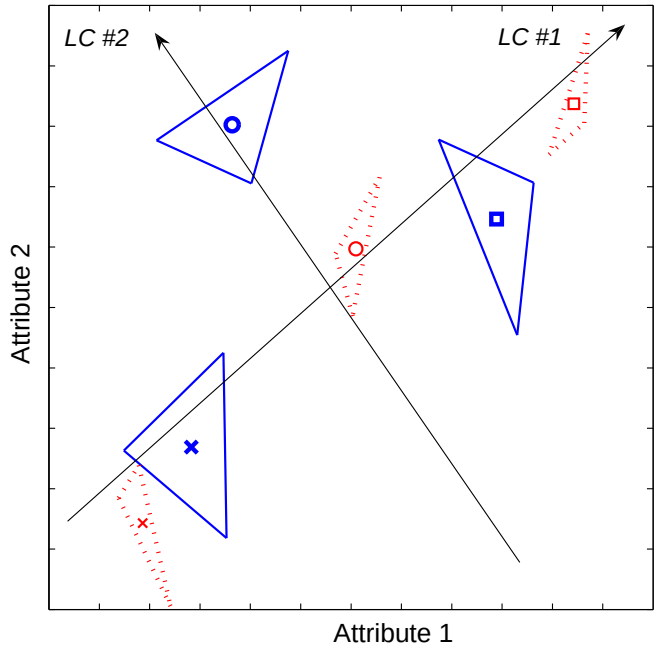


Figure 4.1: Illustration of panelist performance with a simple simulated sensory data set comprising two attributes, three products shown with a different symbol (cross, circle and square) and two assessors S and D shown with a solid blue line and a dotted red line respectively (adapted from Schlich *et al.*, 2004).

latent components LC #1 and LC #2 respectively. In more a realistic case, the sensory profile would comprise a larger number of attributes and a multivariate data analysis would be needed to represent the products on a small set of latent components. Nevertheless, displaying a two-dimensional map of this latent sensory space would give a plot similar to the one presented in Figure 4.1 but with coordinate axes representing two latent variables instead of the two manifest variables displayed in the present example.

Multivariate analysis aims at comparing product positions in the latent sensory space and Schlich *et al.* (2004) makes the difference between two measures to assess product discrimination: strength defined as the distance between products in the sensory map and complexity defined as the dimensionality of the sensory space. Looking at the two-dimensional product configuration of the assessors S and D in Figure 4.1, it appears that the sensory description of the assessor D is one-dimensional (along the latent component LC #1) whereas the description of assessor S is two-dimensional. This indicates a larger correlation between the two attributes for the assessor D. The distance between symbols is also larger for this assessor, which highlights a better discrimination along the first latent component. However, this assessor appears not to be discriminant along the second latent component (LC #2). On the contrary, the assessor S shows some discrimination along this second component despite his low repeatability seen from the large surface of the blue triangles. The multivariate approach gives therefore another perspective to the issue of disagreement in this simulated example.

4.4.4 A review of statistical methods for evaluating consensus panel performance

The practical example considered above illustrated the two complementary views to panel performance evaluation. The univariate approach offers a detailed perspective to the use of the descriptors by the consensus panel whereas the multivariate approach offers a more global representation of the product space on a smaller set of latent dimensions from which information about the use of the consensus vocabulary can also be highlighted. A review of the methods found in the literature for CV panel performance monitoring will now be presented for these two categories (these two categories).

Univariate approaches to panel performance evaluation

Univariate approaches to panel performance evaluation are relatively easy to understand as they map directly to attribute scales. A sensory score is the result of a complex process involving physiological and cognitive aspects. The attribute grading made by an assessor on a linear continuous scale is the last step of this measurement process and it can be affected by individual variations relating to scale usage. These variations can be very large but have usually no relation to product characteristics. They are often considered as nuisance effects by the experimenter and approaches to model these individual differences have been developed in sensometrics. Brockhoff (2003) for instance described four basic assessor differences that can be encountered on a univariate scale: level differences when assessors use different parts of the scale, scaling differences when assessors use different amount of the scale, variability when the precision of assessors differ and disagreement which is the non-linear individual variation not attributable to the three other scaling differences.

Panel monitoring made at an individual attribute level can provide very detailed information about the performance of the assessors. For example, Rossi (2001) employed two measures relating to the repeatability and disagreement criteria defined in the previous section. These measures are computed per product, per assessor and per attribute (the second criterion is referred to as 'reproducibility' in the original paper) and are based on a statistical model defined by Mandel (1991). This approach results in a large set of descriptive statistics, which Rossi summarized with several graphical techniques. Another statistical method based on classical reliability theory was proposed by Bi (2003) to compute the same criteria. However, these two statistics are only computed at a more global level on an attribute basis (for the full panel) and on an assessor basis (for the full set of attributes), which yields only two graphical displays.

Similar visual representations of panel performance have been presented for the reliability and discrimination criteria defined earlier by e.g. Næs and Solheim (1991), Lea *et al.* (1995) and Tomic *et al.* (2007) employing statistical techniques based on the principle of analysis of variance (ANOVA)[4]. These techniques developed at Matforsk

[4]A good introduction to the application of ANOVA to sensory data can be found in (Lea *et al.*, 1997). Note also that performance measures based upon ANOVA methods have been applied to data from the affective domain in the field of audio by Gabrielsson (1979b) and Bech (1992, 1993).

in Norway have been recently implemented in a software package called PanelCheck (Nilsen *et al.*, 2007). Schlich (1994) developed a method for graphical representations of assessor performances based on individual and global ANOVAs per attribute, which also covers the three performance criteria defined above. More details about this approach will be given in Chapter 6. Schlich (1997) proposed later another method based on a similar principle known as 'Control of Assessor Performance' (CAP), which is now part of the SensoBase (2006) online data analysis programme.

Brockhoff (2003) presented a univariate assessor performance method taking into account the four basic assessor differences presented above. This approach relies on a set of statistical models based on ANOVA and includes significance tests for the following assessor effects: differences in variability, presence of disagreement, differences in scaling and differences in sensitivity (defined as a signal-to-noise ratio). Brockhoff's approach often referred to as the 'assessor model' can be considered one of the most thorough univariate tools for assessor performance monitoring and is available as a routine called PANMODEL implemented in the SAS software package. Finally, it can be mentioned that clustering methods or factorial methods have also been used for panel performance assessment of each attribute separately, e.g. Dijksterhuis (1995) and Couronne (1997) presented PCA-based methods to assess the level of disagreement between panelists.

Multivariate approaches to panel performance evaluation

In multivariate approaches to panel performance evaluation, the full set (or a subset) of sensory descriptors is considered which means that a matrix of samples by attributes has to be handled for each assessor. One simple approach to assess panel performance in this case is to measure similarities between such matrices. Ledauphin *et al.* (2006) presented a method based on this idea which consists in computing a weighted average for the panel and deriving an index of agreement for each assessor. The proposed procedure also includes a test of statistical significance for this index. It should be mentioned that several multivariate analysis techniques exploit the same idea. For example the RV coefficient (Robert and Escoufier, 1976) central to the STATIS method (Schlich, 1996) and the Procrustes distance employed in Generalized Procrustes Analsyis (GPA; see Gower, 1975) are also based on the principle of similarity measure between matrices. More details on these two methods will be provided later in this thesis.

Most multivariate methods for panel performance assessment rely on the concept of latent variables discussed in the previous paragraph. Factorial methods are commonly used in sensory data analysis to represent the products in a space of reduced dimensionality. These techniques offer various means to assess individual variability around the consensus usually in the form of visual representations on the sensory map, for example with ellipses of confidence. Husson *et al.* (2005, 2007) presented an assessor bootstrapping technique to assess the variability of the panel for each product and each attribute. Brockhoff (2001) considered using multivariate analysis of variance (MANOVA) and the associated canonical variate analysis (CVA) to account for measurement error in this type of analysis. Monrozier and Danzart (2001) also proposed an assessor re-sampling technique to represent variability around product means in

PCA and MANOVA-CVA models. The use of partial least squares regression models has also been reported by Thybo and Martens (2000) who measured signal to noise ratios (i.e. a ratio of the systematic between-object variation and the residual noise) for attributes, assessors and products with the aim of highlighting problems of assessor disagreement and attribute discrimination.

Another group of multivariate analysis referred to as three-way methods can also be employed to model assessor variability at a more complex level. GPA for example has been applied to consensus vocabulary profiling to test assessor agreement on the interpretation of attributes by allowing certain types of transformations to the individual sensory profiles. Other approaches include the Tucker-1 model described e.g. in (Dahl *et al.*, 2008). Qannari *et al.* (1995) and more recently Bro *et al.* (2008) proposed a hierarchy of three-way models covering methods such as PCA, Tucker-1 and PARAFAC, which allows to characterize the performance of assessors in terms of latent sensory dimensions.

4.5 Summary

The overview of CV methods presented in the present chapter highlighted the following principles of this descriptive analysis method: 1) employing a panel of trained assessors rather than a single expert, 2) developing an objective description of products based exclusively on sensory properties and 3) performing quantitative evaluations with reliable scales. Five important techniques of the class of CV methods were then presented and discussed, i.e. Flavor Profile, Texture Profile, QDA, Spectrum and Quantitative Flavor Profiling. From this review, three steps commonly found in the development of a consensus vocabulary by a sensory panel were identified as follows: 1) the panel selection, 2) the consensus vocabulary generation and 3) the panel training. Finally, the discussion on the concept of 'consensus' in a vocabulary development and the review of desired characteristics for a consensus vocabulary illustrated the important aspects to consider for the practical implementation of this challenging methodology.

The second part of this chapter addressed the issue of consensus vocabulary quality through the use of panel performance measurement. First, a discussion on the general concepts of precision and accuracy highlighted the partial applicability of these measures to sensory analysis. Then, the three criteria of repeatability, agreement and discrimination were defined and their interpretation was illustrated from the univariate and multivariate point of views. Finally, a literature review was presented on existing statistical tools for panel performance monitoring which highlighted the numerous approaches available for this purpose.

Chapter 5

Development of a consensus vocabulary for headphone sound perception

5.1 Introduction

The overview of headphone sound presented in Chapter 3 gave an illustration of the diversity in headphone usage and applications. The examples discussed earlier include the headphone replay of conventional stereo material or binaural recordings and the virtual loudspeaker reproduction of stereo or 5.1 audio material over headphones. The specificities of headphone sound perception were also reviewed and a lack of formal terminology to describe this particular perceptual domain was highlighted. The current thesis work addressed this latter issue by developing a set of attributes to describe headphone sound perception using the consensus vocabulary (CV) approach.

Following the overview of CV methods presented in Chapter 4, a practical application of this type of descriptive analysis to the perceptual domain of headphone sound is reported in the present chapter. In Section 2, the stimulus selection, the procedure employed for the consensus vocabulary development and the resulting terminology are presented in details. In Section 3, two additional steps of this experimental work are reported focusing respectively on the development of a set of sound exemplars for attribute anchoring and the comparison of two attribute rating methods. In Section 4, the application of the resulting set of attributes is illustrated with a simple set of headphone sound stimuli. Finally, the implementation and outcome of this consensus vocabulary development are discussed in Section 5.

5.2 Consensus vocabulary development

The consensus vocabulary development reported in this section conforms with the general principles described in the previous chapter and its implementation can be summarized as follows. A wide range of stimuli was selected to cover the perceptual space under study and was presented to a panel of 12 listeners who developed a set of consensual sensory descriptors with associated definitions and word anchors, under the

supervision of a panel leader who was not involved in the elicitation process. An effort was also made to build a set of audio exemplars illustrating each sensory descriptor and to define an intensity value on their associated scale. And finally, the different attribute rating tasks organized in this experiment were performed individually by each assessor.

It appears from this short overview that the methodology applied for the current study mixes characteristics from some of the CV techniques presented earlier. For example, the nature of the sensory panel follows the QDA approach by using a panel of non technically-biased assessors while the development of a set of audio exemplars for absolute intensity scaling resembles the QFP approach. The present method would therefore fit best the class of generic (consensus) descriptive analysis described by Einstein (1991) or Murray *et al.* (2001).

In this section, the development of a representative set of headphone sound stimuli is reviewed first. Some background on the methodology is then provided and the experimental steps of this vocabulary development work are described. Finally, the agreed terminology is presented.

5.2.1 Selection of a stimuli set

The first step of this study was to select a set of audio samples for the listening experiment. Building a representative set of stimuli was considered critical to the success of this work which aims at exploring the perception of sound reproduced over headphones in a broad sense. Covering all the perceptual aspects relevant to headphone sound is especially important when a descriptive analysis technique is applied because only the characteristics that are present in the sample set can be elicited by the listeners. Therefore, a systematic classification of the commonly encountered headphone sound scenarios was developed in this study to facilitate the creation of an appropriate database of audio samples for the vocabulary development phase.

The classification developed for this task presented in Table 5.1 includes three main parameters defined respectively as the sound material format, the sound reproduction format and the sound reproduction technique. The format of the original sound material shown in the first column is not restricted in headphone applications and a wide selection was therefore applied for this study including the mono, stereo, multichannel[1] and binaural formats. On the contrary, the reproduction format defining the number of channels employed for the replay is naturally limited to two in the case of headphones, which leaves only the mono and stereo options as shown in the second column of the table. The third column in Table 5.1 specifies the approach employed for the reproduction of the source material. For example, the stereo reproduction of a mono sound source can be made by a direct diotic or dichotic replay or can follow some HRTF or BRIR filtering process. Different types of spatial enhancement techniques can also be applied for the (stereo) reproduction of stereo or multichannel sound material over headphones as illustrated in this table. Alternatively, binaural

[1]From the various multichannel audio formats described in Chapter 3, only the 5.1 format was selected as it was considered sufficiently illustrative for the present study.

material recorded directly with an artificial head can be replayed dichotically over headphones.

The fourth column in Table 5.1 gives more details about the signal processing applied to the sound material. The perceived room effect represented by a sound stimulus is an important aspect of this classification and this characteristic can be captured directly in the original sound material recorded, for example in different acoustic environments, or/and it can be added to this audio material through various types of audio processing. In the latter case, the sound format (column 1 of Table 5.1) and the sound reproduction format (column 2) define the scope of the applied signal processing. For example, only a mono reverberation can be applied in the mono reproduction case but the stereo reproduction case brings more options in terms of room effect simulation. The mono-format/stereo-reproduction scenario offers additional levels of complexity such as the stereo panning technique selected and the parameters applied for headphone positional 3D audio processing, e.g. the HRTF filter type or the sound source direction and distance. Similarly, different spatial enhancement systems and settings can be considered for the reproduction of stereo or multichannel audio material over headphones, as described in Lorho *et al.* (2002) and Lorho and Zacharov (2004).

This classification was used as a basis for the stimulus generation task. Different types of sound source material were selected including male and female speech samples, musical instruments, more complex music samples, natural sound recordings, movie sound samples and synthetic sound samples. This audio material included also different types of room acoustics with recordings made in an anechoic room, a standard listening room and a reverberation chamber. For each category listed in Table 5.1, a collection of sound source material was selected. Applying a full combination of the sound material types, sound formats, reproduction formats, reproduction techniques and signal processing approaches was of course not possible due to the incompatibility between some of the categories but even when selecting all the practically feasible combinations, the size of the resulting database would be very large. Therefore, only an incomplete combination of the different factors was selected to create a database of manageable size, i.e. about 200 audio samples.

Following the development of this headphone sound database, a final set of stimuli comprising 50 audio samples was chosen based on an informal screening procedure. This selection included samples that formed a first layer illustrating large perceptual differences, i.e. audio samples selected from different sound reproduction categories and audio material. A second layer illustrating moderate to small perceptual differences was also considered by selecting pairs of stimuli from a given sub-category. For example, the same source material was included for two different types of stereo panning or for two different options of the same virtual 5.1 system. This process resulted in a generic set of stimuli considered representative of the important perceptual characteristics of headphone sound reproduction, including lateralization, externalization (i.e., inside-the-head and out-of-head locatedness), localization and aspects relating to space and timbre perception.

Sound format	Reproduction format	Reproduction technique	Characteristics of the applied sound processing
Mono	Mono	Monotic (direct)	Dry
			Added mono reverberation (different levels)
	Stereo	Diotic (direct)	Dry
			Added mono reverberation (different levels)
		Dichotic (direct)	Added stereo reverberation (different levels and types)
		Stereo panning	Amplitude panning (soft to hard levels of panning)
			Amplitude and time based panning (different levels of panning)
		Head-Related Transfer Functions (HRTFs)	Simplified/modelled HRTF Model (different sound source directions)
			Measured HRTFs (different sound source directions, different HRTF filter types)
			HRTF + artificial reverberation (mono or stereo)
		Binaural Room Impulse Responses (BRIRs)	Modelled BRIRs (different types)
			Measured BRIRs (different rooms)
			Troncated BRIRs (different lengths of room response)
		Mono enhancement	Mono-to-3D processing
Stereo	Mono (downmix)	Monotic / Diotic	Case similar to the mono source format
	Stereo	Dichotic (direct)	Different stereo formats, e.g. 90º cardioid and Blumlein
		'Soft' stereo enhancement	e.g. stereo widening algorithm (Kirkeby, 2002)
		HRTF-based stereo enhancement	Different systems and settings[1]
		BRIR-based stereo enhancement	Different systems and settings[1]
Multichannel (5.1)	Stereo	'Soft' virtual 5.1 processing	See Lorho (2006) for details
		HRTF-based virtual 5.1 processing	Different systems and settings[1]
		BRIR-based virtual 5.1 processing	Different systems and settings[1]
Binaural rec. material	Binaural reproduction	Dichotic (direct)	Different types of binaural recordings and audio environments

[1] See Lorho (2006) for details

Table 5.1: Classification of headphone sound scenarios used as a basis for the creation of a database of sound samples for this vocabulary development work.

5.2.2 Vocabulary development methodology

Once this generic set of stimuli was created, the consensus vocabulary development could be initiated following the generic steps presented in Chapter 4 (Section 4.3). Before going through the implementation details of these steps, some background on the selected descriptive analysis methodology will be presented below.

The perceptual study presented in this chapter was inspired from the Audio Descriptive Analysis & Mapping (ADAM) procedure developed by Zacharov and Koivuniemi (2001b) and Mattila (2001a). The ADAM technique was proposed as an experimental means of unraveling the underlying structure of a perceptual domain in the field of audio. This procedure requires two separate listening experiments to be run on the stimulus set under study, i.e., a preference scaling task on the one side and a verbal descriptive analysis using a consensus vocabulary approach followed by an attribute rating task on the other side. The last step of the ADAM method consists in applying an external preference mapping, which consists in a combined analysis of the two data sets resulting from this set of experiments. The aim of this mapping technique is to explain the global measure obtained from the preference test in terms of the more detailed sensory characteristics elicited from the descriptive analysis. A predictive model of preference can also be built based on the attribute rating data[2].

The vocabulary development performed in the present study followed the descriptive analysis part of the ADAM methodology but some refinements to the original process have been proposed by the author with the aim of reducing the time spent for the vocabulary development. The previous ADAM experiments required a considerable amount of time to develop the set of descriptors, i.e., 60 hours for the study of Mattila (2001a) and 30 hours for the study of Zacharov and Koivuniemi (2001b). This was considered a constraining factor to the usage of this approach routinely and several directions were therefore sought to increase the speed of the procedure. First, it was realized that too much effort and time was spent for the initial word elicitation. In practice, the individual elicitation can be reduced significantly without major impact on the final outcome of the whole process because this step is only a preparation for the more thorough consensus elicitation task. Then, a second modification of the procedure was considered with the introduction of an iterative step during the last step of the consensus vocabulary development. This idea consisted in combining group discussion sessions with the whole panel and practice sessions in which each assessor can experiment individually with the attribute scales to ensure a better alignment of the sensory concepts by the panel both at a qualitative level and a quantitative level, as discussed in Section 4.3.2. These methodological refinements allowed a significant reduction of the vocabulary development process duration to a total of 20 hours.

5.2.3 Presentation of the vocabulary development steps

The descriptive analysis experiment performed in this study can be loosely separated into two successive phases including first the vocabulary development work made by the panel and then the set of attribute rating tests made individually by the assessors.

[2]An overview of preference mapping can be found in Chapter 2 (Section 2.2.4).

The first phase was implemented within a relatively short period of two weeks due to practical reasons relating to the availability of assessors for the large panel discussions. The second phase to be reported in Sections 3 and 4 of this chapter was spread over a slightly longer period of time, partly because of the experimental preparation needed between the test sessions. The verbal elicitation was carried out in Finnish as this was the native language of the assessors and the panel leader. In practice, all the individual listening sessions were run in separate listening booths (Kylliäinen *et al.*, 2003) and all panel discussions were organized in a large listening room equipped with 12 pairs of HD580 Sennheiser headphones. The loudness of all audio stimuli was aligned beforehand and replayed at a comfortable listening level of 25 sones. The GuineaPig2 listening test system (Hynninen and Zacharov, 1999) was employed for the stimulus replay in all the listening tests included in this study.

The vocabulary development part of this experiment was run in 20 hours and was organized in four steps including a familiarization of the panel to the stimulus set, an individual elicitation, a set of small panel discussions, and a set of large panel discussions, as illustrated in the diagram shown in Figure 5.1. Note that the selection of an appropriate sensory panel is a prerequisite for this type of sensory analysis experiment, as highlighted in Chapter 3. A group of 12 listeners was chosen from the Nokia Research Center (NRC) listening panel for the present study. These subjects can be considered as 'selected assessors' following the classification presented in Chapter 2 because they have been screened through the Generalised Listener Selection (GLS) procedure (Isherwood *et al.*, 2003) for their auditory discrimination skill, reliability and repeatability in judgments and their verbalization skills. In addition, all these assessors showed good discrimination skills in the task of the GLS screening program relating specifically to the perception of spatial sound over headphones[3].

For the stimulus familiarization step, a listening test session of 30 minutes to one hour was organized in individual listening booths to accustom the assessors to the set of stimuli under study. The task of the subjects was simply to listen to each of the audio samples in a sequential way and they were instructed to concentrate on the perceptual characteristics of these stimuli. This preliminary step was considered especially important because of the complex nature and diversity of the audio samples considered in the present study.

The individual elicitation followed immediately the familiarization step and was split into two separate listening test sessions. Assessors performed an absolute word elicitation during the first session which lasted about an hour. In practice, a small subset of 16 stimuli was presented using a single stimulus paradigm and for each sound sample listeners were required to write down words describing their impression on any aspect of the reproduced sound. During the second word elicitation, assessors performed a differential word elicitation on the remaining audio samples split in two subsets of 14 and 20 stimuli respectively. This second listening test session lasted about two hours with a short break. In practice, all the samples of a subset

[3]The stimulus set employed for this discrimination task will be described more thoroughly in Section 4 of the current chapter.

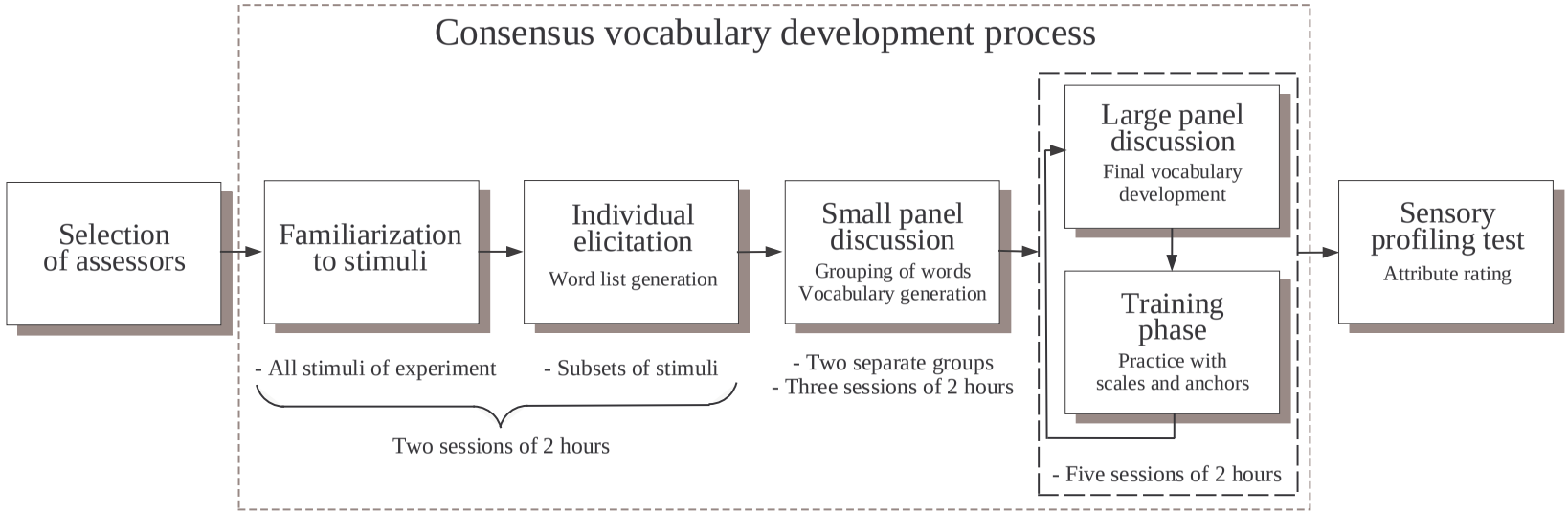


Figure 5.1: Illustration of the steps included in the consensus vocabulary development procedure employed for this experiment. The implementation of process required a total time of 20 hours split into 4 hours of individual work, 6 hours of small panel work and 10 hours of large panel work.

were presented with a multiple stimulus paradigm and listeners were asked to focus on differences between the two samples of each pair of stimuli relating to the same audio material as defined in the previous section. The stimulus familiarization step and the individual elicitation took a total 4 hours as illustrated in Figure 5.1 and the outcome at this stage of the procedure was a set of about 500 words generated by the group of 12 assessors, which is considerably less than the 26000 and 1400 words elicited respectively in Mattila (2001a) and Zacharov and Koivuniemi (2001b).

The individual elicitation work can be considered as a preparation to the more essential step of consensus vocabulary development in which the panel of assessors develops a common set of sensory descriptors under the supervision of the panel leader. this part of the procedure was split into two separate phases as illustrated in Figure 5.1. First, small panel discussions were run in two separate groups of 6 assessors as this was considered a more efficient approach both in terms of verbal elicitation outcome and practical implementation. The small panels started by discussing and sorting their individual set of words from the previous elicitation phase and agreed rapidly on an initial set of descriptors. In three sessions of two hours, the two panels created a set of 18 and 16 descriptors respectively.

Large panel discussions were then organized in which all 12 assessors met together and finalized the vocabulary generation based on the sensory descriptors developed during the small panel sessions. Three sessions of two hours were used by the panel to compare the two sets of descriptors and consensually agree upon a final vocabulary to describe the set of headphone stimuli under study. They also created a suitable set of word anchors, i.e. a verbal term defining a certain point on a graphic scale (Zielinski *et al.*, 2008) for the end points and the middle point of each attribute scale and they selected illustrative audio samples from the stimulus set for each of the attributes. A short definition was also proposed by the panel for all the descriptors and their explanation were thoroughly discussed to ensure a common understanding by all the assessors. It should be noted that a short phase of panel training was also considered in this vocabulary development process. This ultimate step of the consensus work was included in the two last sessions of the fourth block shown in Figure 5.1 and consisted in two rounds of alternate group discussions and individual experimentation with attribute scales. This iterative process allowed to improve the alignment of the sensory concepts between the assessors to some extent through feedback given by the panel leader on attribute scale usage.

5.2.4 Presentation of the agreed vocabulary

A detailed description of the terminology resulting from the consensus vocabulary development work performed in the current study is provided next. The vocabulary agreed on by the panel comprises a set of sixteen sensory descriptors with associated definitions and word anchors. The original name of the descriptors are given below in Finnish and an English translation is proposed for the attribute name, the definition and the set of word anchors developed by the panel. The listeners agreed on a division into three groups of descriptive scales relating respectively to localization, space and timbre. Attributes in the localization group describe simple geometrical associations

between the perceived sounds and the listener. The emphasis of this group is on definability, i.e. the possibility for the listener to define a certain aspect of the sound source(s) he/she hears. Attributes of the second group relate to the space perceived by the listeners in the audio samples and the attributes of the third group relate broadly to timbre aspects of the sound samples.

Attributes relating to localization

- **Sense of distance** (*Etäisyyden tunne*): this attribute describes how well the distance between the sound source(s) and the listener can be defined. Word anchors: not definable (0), somewhat definable (5), well definable (10).
- **Sense of direction** (*Suunnattavuus*): this attribute describes how well the direction of the sound source(s) can be defined. Word anchors: not definable (0), somewhat definable (5), well definable (10).
- **Sense of movement** (*Liikkeen tunne*): this attribute describes how well the movement of the sound source(s) can be defined. Word anchors: not definable (0), somewhat definable (5), well definable (10).
- **Ratio of localizability** (*Paikallistettavien osuus*): the term *localizability* describes how well the direction and the distance of a sound source(s) can be defined. The attribute *ratio of localizability* describes how many sound events can be localized from those present in the audio sample. Word anchors: none (0), some (5), all (10).

Attributes relating to space

- **Quality of echo** (*Kaiun laatu*): this attribute describes how well the echoes relate to their sound source(s) in a qualitative way. Word anchors: unpleasant (-5), no echo (0), pleasant (+5).
- **Amount of echo** (*Kaiun määrä*): this attribute describes how the listener experiences the amount of echo in relation to the sound sources. Word anchors: no echo (0), adequate echo (5), excessive echo (10).
- **Sense of space** (*Tilan määriteltävyys*): this attribute describes how well the space represented in the audio sample can be defined. Word anchors: not definable (0), somewhat definable (5), well definable (10).
- **Balance of space** (*Tilan tasapaino*): this attribute relates to the space represented by the audio sample in relation to the listener's inner reference. A negative value means that the space is weighted in some direction. If no space is perceived, the space is out of balance. Word anchors: out of balance (0), somewhat balanced (5), well balanced (10).
- **Broadness** (*Laajuus*): this attribute describes the perceived extent of the soundscape relative to the listener's head. Word anchors: inside head (0), close by (5), broad (10).

Attributes relating to timbre

- **Separability** (*Eroteltavuus*): this attribute describes how well the sound events can be separated out in the audio sample. Word anchors: none (0), some (5), all (10).

- **Tone Color** (*Äänen sävy*): this attribute describes the spectral content of the audio sample. Word anchors: lower-sound emphasis (-5), cannot say (0), higher-sound emphasis (+5).

- **Richness** (*Värikkyys*): this attribute describes how rich and nuanced the audio sample is overall, and relates to a combination of harmonics and dynamics perceived in the sample. Word anchors: flat (-5), neutral (0), rich (+5).

- **Distortion** (*Vääristyneisyys*): this attribute describes the possible metallic, machine-like, electrical-like artifacts in the audio sample. Word anchors: distorted (0), somewhat distorted (5), not distorted (10).

- **Disruption** (*Häirioisyys*): this attribute describes how much hiss, snap/crackle/pop is perceived in the audio sample. Word anchors: disrupted (0), somewhat disrupted (5), not disrupted (10).

- **Clarity** (*Selkeys*): this attribute describes if the sound sample appears clear of muffled, for example if the sound source is perceived as covered by something. Word anchors: muffled (0), clear (10).

- **Balance of Sounds** (*Äänien voimakkuuksien tasapaino*): this attribute describes the possible difference in loudness between the sound sources present in the audio sample. The sound sample is well balanced if it contains only one sound source. Word anchors: out of balance (0), somewhat balanced (5), well balanced (10).

It appears from the numerical value associated to the different sets of word anchors that most of the attributes created during the vocabulary development are unipolar in the sense that they measure an intensity in a single direction, e.g. *Broadness* increases from 0 (*Inside the head*) to 10 (*Broad*). However, three attributes of this vocabulary can be considered bipolar as they have been anchored by the panel with a center point and two opposite directions. These attributes are *Tone Color* going from -5 (*Lower-sound emphasis*) to +5 (*Higher-sound emphasis*) with the middle-point (*Cannot say*) at 0; *Richness* going from -5 (*Flat*) to +5 (*Rich*) with the middle-point (*Neutral*) at 0; and *Quality of echo* going from -5 (*Unpleasant*) to +5 (*Pleasant*) with the middle-point (*No echo*) at 0[4]. It can also be noted from this list that the two attributes *Distortion* and *Disruption* describing different artifacts of the audio samples were associated with a 'negative concept' by the listening panel and have therefore their polarity inverted, i.e., a very 'distorted' sound would be given a low score on that scale.

[4]The validity of the attribute *Quality of echo* as a descriptive term can be questioned because of the hedonic nature of its positive and negative word anchors and the integrative character of the term *Quality*.

5.3 Additional steps of the consensus vocabulary development work

Following the phase of consensus vocabulary development described above, the sensory panel went through four additional experimental tasks including 1) the creation of a set of sound exemplars with associated intensity anchors for each descriptor, 2) a small listening experiment to compare two attribute rating approaches, 3) an attribute rating test performed on a simple stimulus set, and 4) a large experiment to assess several sets of spatial enhancement systems for music reproduction over headphones. Details about the first and second tasks are provided in the present section while the third task is covered in Section 4 of this chapter. Finally, the perceptual evaluation of spatial enhancement systems will be covered in Chapter 6.

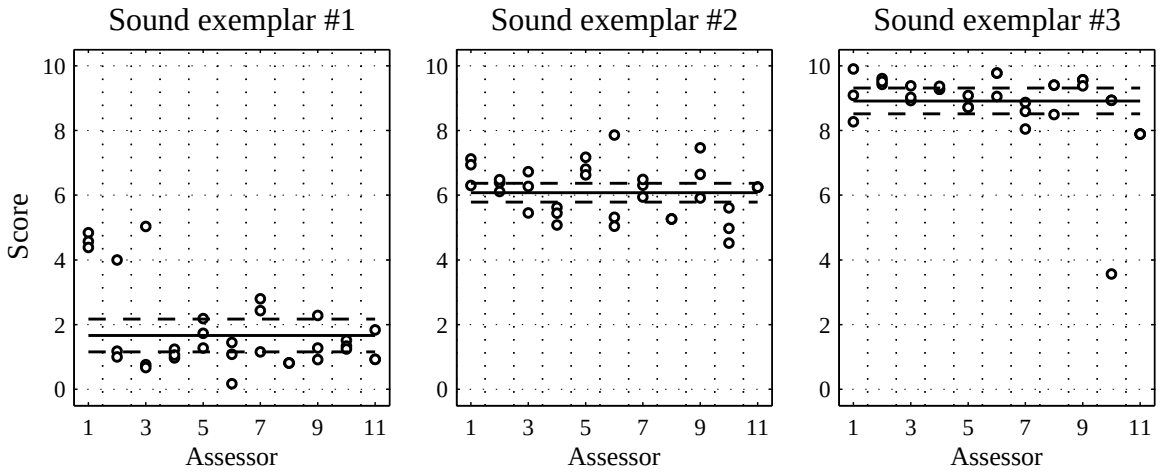
5.3.1 Selection of sound exemplars for attribute anchoring

The importance of physical references to illustrate sensory concepts during a vocabulary development and the advantages of selecting absolute reference points for attribute scale anchoring were discussed in the previous chapter (Section 4.3.2). Following these recommendations from the literature (see e.g. Rainey (1986) and Munoz and Civille (1998)), an effort was made during this experiment to create a set of audio exemplars illustrating the different descriptors of the headphone sound vocabulary. As described earlier, one of the tasks of the listening panel during the last step of the vocabulary development process was to select illustrative audio samples from the headphone sound database for each of the attributes. However, due to the limited amount of time allocated for this task, the selection was only partially completed by the panel and the set of sound exemplars had to be augmented and refined by the experimenter. It was therefore decided to run a formal listening test with the panel to identify the best set of sound exemplars and to quantify them on their respective attribute scale. In practice, a series of candidate audio exemplars was presented randomly in triplicates amongst other stimuli for each attribute using a single stimulus test paradigm and each assessor had to score these stimuli on the associated attribute scale.

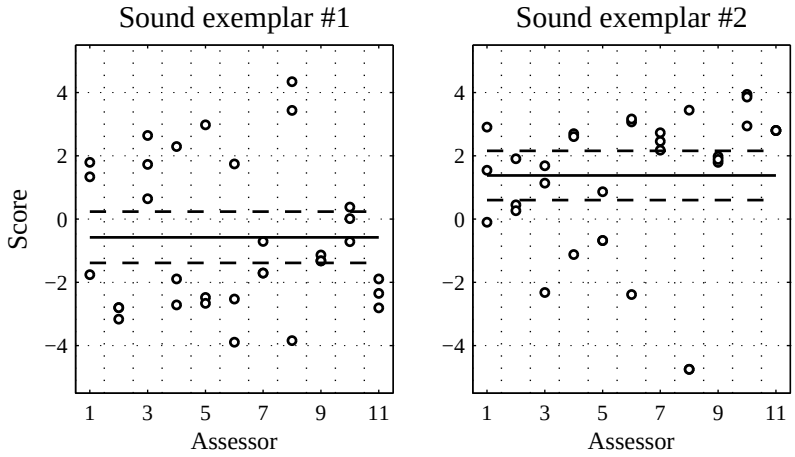
The data gathered from this experiment was analyzed visually with the help of scatter plots as illustrated for two representative attributes in Figure 5.2. The intensity anchoring obtained for the attribute *Amount of Echo* shown in Figure 5.2(a) can be considered successful because the three sound exemplars cover well the attribute scale with a mean value over repetitions and assessors of 1.6, 6.1 and 8.9 respectively and also because the standard deviation of these mean values is relatively low[5]. On the contrary, the result of the attribute *Richness* shown in Figure 5.2(b) was less conclusive. Only two sound exemplars were considered for this attribute and they appear

[5]It can be noted from Figure 5.2(a) that Assessor #1 was clearly in disagreement with the rest of the panel for the sound exemplar #1 (in the left pane). Ignoring this outlier, the panel average score for this sound exemplar was 1.4, which is the reason why the intensity 1 was finally associated to this sample as shown in Table 5.2.

to cover the attribute scale poorly with an average over repetitions and assessors of -0.6 and 1.4 respectively. A more serious problem with this attribute concerns the large standard deviation visible for the two audio exemplars (dashed lines on the left and right panes of Figure 5.2(b)), which indicates a clear lack of agreement between the assessors. In this situation, the anchoring of the sound exemplars by a numerical intensity corresponding to the panel mean was considered inappropriate. Instead, a negative sign and a positive sign were used to provide a simple qualitative illustration of the attribute polarity.



(a) Scatter plot of intensity values obtained for the three audio exemplars of the attribute *Amount of echo*.



(b) Scatter plot of intensity values obtained for the two audio exemplars of the attribute *Richness*.

Figure 5.2: Two examples of sound exemplar assessment made by the panel. Each circle represents one of the three scores given by an assessor and the horizontal lines represent the mean over repetitions and assessors (solid lines) and ± standard deviation (dashed lines) for a given sound exemplar.

The sound exemplars selected from this experiment and their associated anchoring intensity values are presented for each attribute in Table 5.2. It appears from this table that only two or three attributes are illustrated by three sound exemplars spanning the whole attribute scale while eight attributes are represented by only two intensity anchors. Additionally, the sound exemplars of six attributes were not considered consensual, as indicated by the − and + symbols shown in the table. These problematic attributes indicate that either the sound exemplar selection was sub-optimal or additional work would be needed with the panel to improve the level of sensory concept alignment.

		Intensity of sound exemplars		
Group	**Attribute**	**Exemplar 1**	**Exemplar 2**	**Exemplar 3**
Localization	Sense of distance	−		+
	Sense of direction	1		9
	Sense of movement	−		+
	Ratio of localizability	4		9
Space	Quality of echo	−		+
	Amount of echo	1	6	9
	Sense of space	−		+
	Balance of space	1		9
	Broadness	1	5	9
Timbre	Separability	−		+
	Tone Colour	-3	0	3
	Richness	−		+
	Distortion	2		9
	Disruption	1		9
	Clarity	1		9
	Balance of Sounds	2		9

Table 5.2: Set of sound exemplars with associated anchoring intensity values resulting from the attribute test reported in this section.

5.3.2 Selection of an attribute rating method

During the last stage of the vocabulary development process, the selection of an appropriate attribute rating technique became an issue both in terms of methodology and practical implementation. In principle, several experimental paradigms can be adopted to profile a set of audio stimuli under study, that is, obtaining for each stimulus an intensity score on the different sensory descriptors. The three following ap-

proaches were investigated in this work: 1) the 'single stimulus - single attribute' presentation method, 2) the 'single stimulus - multiple attributes' presentation method, and 3) the 'multiple stimulus - single attribute' presentation method.

The first approach has been utilized for example by Mattila (2001b) in the field of speech perception but the second approach is applied more frequently, especially in the field of food science where it is referred to as the 'serial monadic' paradigm[6]. The third approach is also known as 'attribute-by-attribute' paradigm or 'simultaneous multiple presentation method' and has been compared to the serial monadic in several publications (Mazzucchelli and Guinard, 1999; Ishii *et al.*, 2007, 2008). Ishii *et al.* (2008) described some fundamental differences between the serial monadic and attribute-by-attribute paradigm in terms of cognitive strategy. They associate the former approach to an absolute cognitive process and the latter one to a relative process in which assessors compare the stimuli under evaluation. These authors argue that the attribute-by-attribute protocol is better suited for untrained judges while the serial monadic protocol is more appropriate when intensity exemplars have been defined and learned by the assessors.

The 'single stimulus - single attribute' presentation method is statistically sound because it ensures an independent grading when the presentation order is randomized for both the stimuli and the attributes. It was however not selected for the current study due to technical limitations with the user interface presentation structure of the GP2 test software. In practice, this approach would have made the test administration too cumbersome for the assessors.

The 'single stimulus - multiple attribute' presentation method breaks the statistical independence of the grading since assessors can consider different scales in parallel and might have a tendency to correlate their attribute scores in a systematic manner. As a matter of fact, a specific order for the evaluation of the different attributes had been specified by the panel during the vocabulary development of this study. This order takes into account the fact that some descriptors are easy to perceive while others require more efforts or a longer exposure to the stimuli. The 'single stimulus - multiple attribute' approach was therefore the first to be considered in this work and it required in practice the presentation of two different windows side by side to get all sixteen attributes on the same screen as shown in Figures 5.4(a) and 5.4(b). It should be noted that the inclusion of sound exemplars was technically unfeasible with the GuineaPig2 listening test system in this test presentation scenario.

The 'multiple stimulus - single attribute' presentation method illustrated in Figure 5.3 was also considered in this work. This comparative evaluation approach enables assessors to better focus on the differences between stimuli although it breaks somehow the statistical independence of the grading. From a practical point of view, the inclusion of sound exemplars was possible with the GuineaPig2 system for this approach (see figure 5.7 for an illustration), which allows in theory to preserve the absolute character of the stimulus evaluation. It should be added however that this experimental paradigm becomes problematic when a large number of stimuli has to be evaluated comparatively.

[6]Note that the generic term 'semantic differential' can also refer to this type of test paradigm.

Comparison of the 'single stimulus - multiple attribute' presentation method and the 'multiple stimulus - single attribute' presentation method

A formal comparison between the 'single stimulus - multiple attribute' presentation method and the 'multiple stimulus - single attribute' presentation method was considered to identify the most appropriate option for the final sensory profiling experiments of this study. A set of four headphone spatial enhancement systems was selected and applied to the same stereo music material for this small listening test. Ten assessors of the headphone panel evaluated these stimuli with each presentation method in a separate session without repetition. The GP2 user interfaces employed for this experiment are shown in Figures 5.3 and 5.4.

To assess the merit of the two presentation methods, a two-way analysis of variance including the factors `System` and `Assessor` was run for each attribute and each test. An example of ANOVA table is shown in Table 5.3 for the attribute *Tone Color*. In each of these tables, the F-ratio obtained for the two factors was exploited for the comparison of the presentation methods. This statistic is a ratio of estimated variances (the mean square) between a factor and the error term. The larger the value of a F-ratio, the more important the differences between levels of the associated factor[7]. In

[7]The absence of replicates in these two data sets implies that the error term employed for computing the F-ratio of the factors `System` and `Assessor` corresponds to the source of variation of the `System` $*$ `Assessor` interaction. More details about the application of ANOVA to univariate sensory data can be found in Section 3 of Chapter 6.

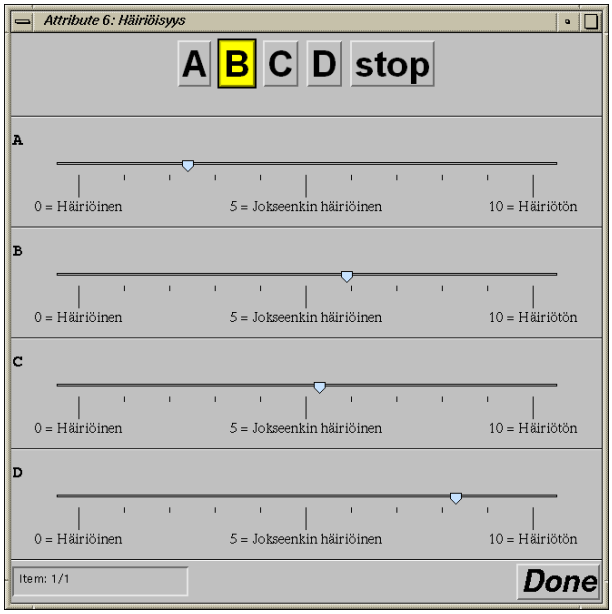
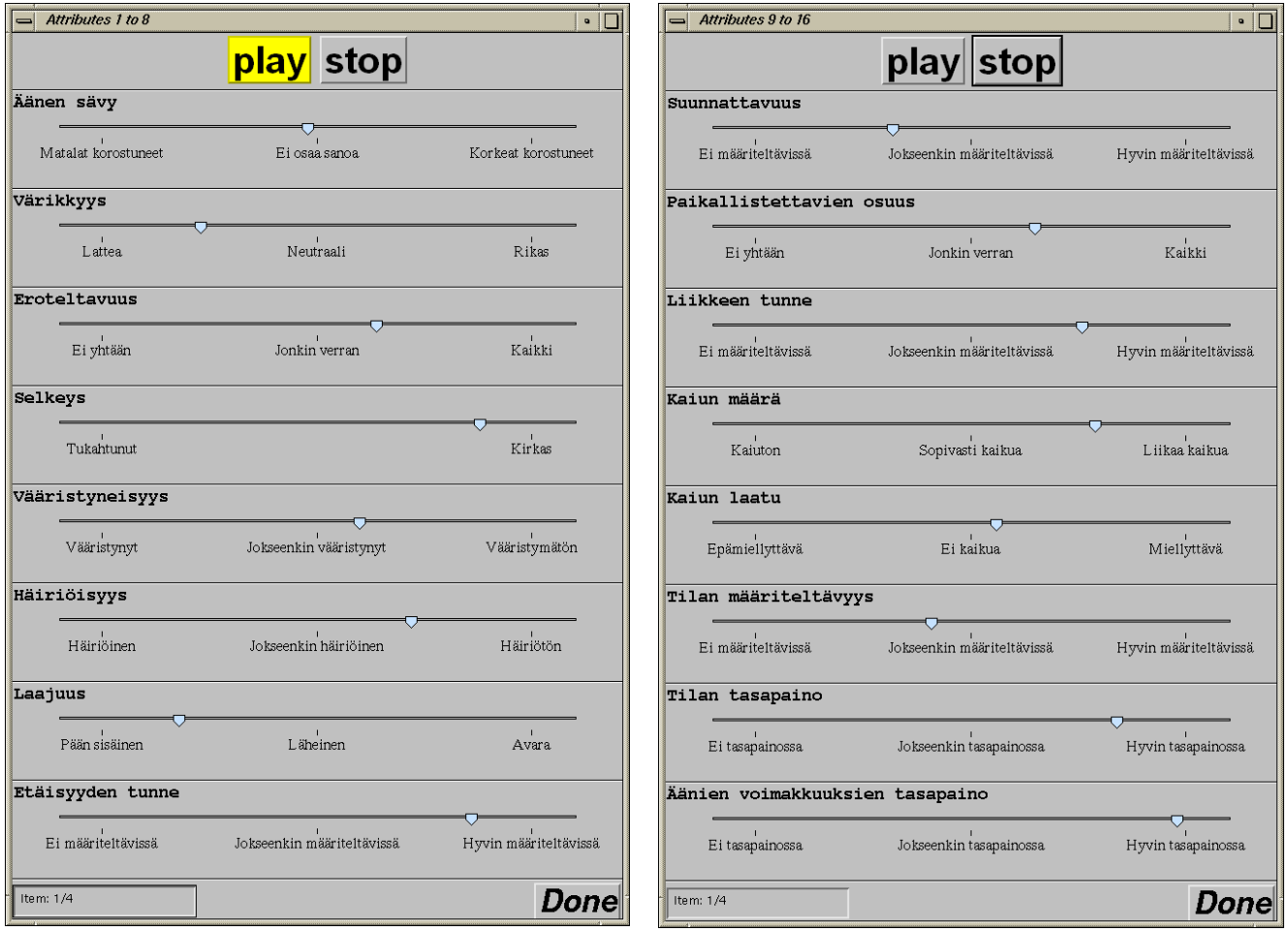


Figure 5.3: Illustration of an attribute rating test using the 'multiple stimulus - single attribute' presentation method. This GP2 user interface window was employed to evaluate the attribute **Disruption** (*Häirioisyys*) in the current experiment.



(a) Attributes 1 to 8

(b) Attributes 9 to 16

Figure 5.4: Illustration of an attribute rating test using the 'single stimulus - multiple attribute' presentation method. The GuineaPig2 user interface windows presented in (a) and (b) were shown simultaneously to the assessors in this experiment (note that the same audio sample could be played from the two windows).

the present case, the `System` F-ratio illustrates the level of panel discrimination between the systems and the `Assessor` F-ratio gives an indication of the differences in scale usage between the assessors.

A scatter plot was produced for each presentation method with the `System` and `Assessor` F-ratios of the 16 attributes on the horizontal and vertical axis respectively, as shown in Figure 5.5. The mean over the 16 attributes of the `System` and `Assessor` F-ratios is also represented by a cross in each plot. It appears from this figure that the overall sample discrimination is higher for the 'multiple stimulus - single attribute' case (18.1 vs. 6.7) while the overall level of differences in scale usage is almost similar for the two presentation methods (3.3 vs. 3.1).

The overall result of this small comparative study and the informal feedback received from the panelists were both favorable to the 'multiple stimulus - single attribute' presentation method. As no further sensory panel training was intended at this stage of the study, it was finally decided to select this approach for the subsequent attribute rating experiments of the study. It should be noted however that this conclusion regarding the relative merit of the two tested presentation methods is limited to the context of the current sensory analysis experiment, i.e. using a panel with the level of training specified above, and should therefore not be generalized.

'Single stimulus – multiple attribute' presentation method
Dependent Variable: *Tone color*

Source	Type III Sum of sq.	df	Mean Square	F	Sig.
Corrected Model	11.255(a)	12	0.938	2.388	0.029
Intercept	0.900	1	0.900	2.291	0.142
Assessor	3.705	9	0.412	**1.048**	0.430
System	7.550	3	2.517	**6.407**	0.002
Error	10.605	27	0.393		
Total	22.760	40			
Corrected Total	21.860	39			

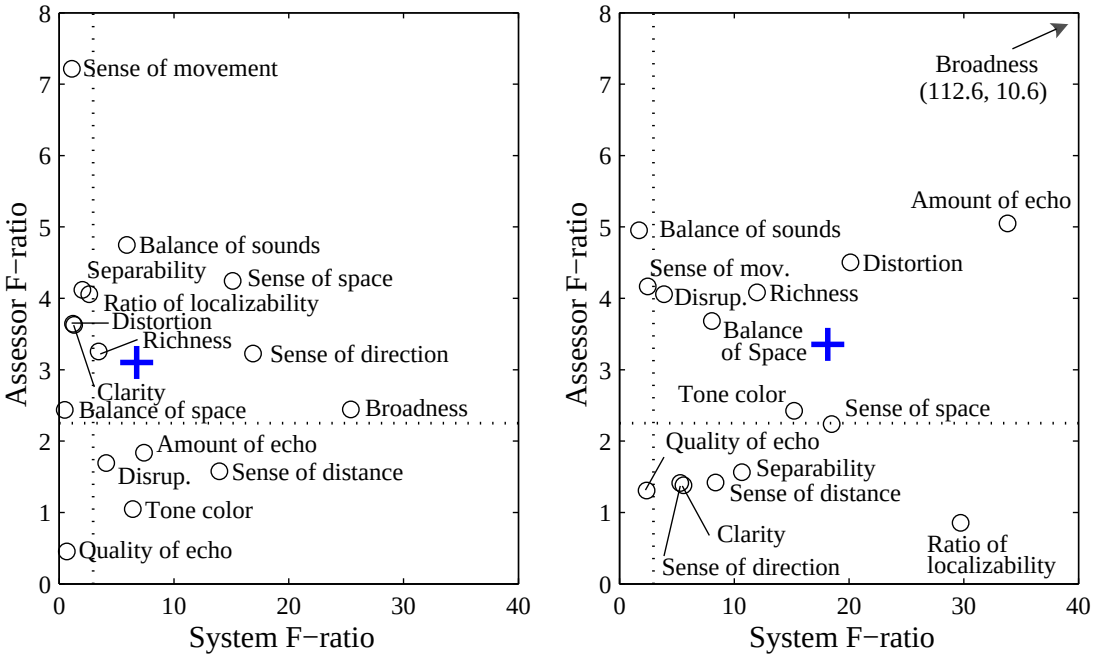
a. R Squared = .515 (Adjusted R Squared = .299)

'Multiple stimulus – single attribute' presentation method
Dependent Variable: *Tone color*

Source	Type III Sum of sq.	df	Mean Square	F	Sig.
Corrected Model	48.100(a)	12	4.008	5.622	0.000
Intercept	1.089	1	1.089	1.527	0.227
Assessor	15.551	9	1.728	**2.423**	0.036
System	32.549	3	10.850	**15.217**	0.000
Error	19.251	27	0.713		
Total	68.440	40			
Corrected Total	67.351	39			

a. R Squared = .714 (Adjusted R Squared = .587)

Table 5.3: Two-way ANOVA table for the attribute *Tone Color* with the 'single stimulus - multiple attribute' and 'multiple stimulus - single attribute' presentation methods.



(a) 'Single stimulus - multiple attribute' presentation method

(b) 'Multiple stimulus - single attribute' presentation method.

Figure 5.5: Scatter plot of *Assessor* and *Sample* F-ratios for the two presentation methods. Comparing in these two graphs the blue cross representing the mean over the 16 attributes of the `System` and `Assessor` F-ratios, the overall sample discrimination can be seen to be higher for the 'multiple stimulus - single attribute' presentation method.

5.4 Application of the consensus vocabulary to a simple stimulus set

5.4.1 Presentation

To illustrate the application of the consensus vocabulary developed in this work, an attribute rating test performed on a simple stimulus set is reported in the present section. The four audio stimuli considered in this sensory evaluation are representative of spatial sound perception over headphones and were developed for the headphone screening test of the Generalized Listener Selection (GLS) procedure (Isherwood *et al.*, 2003). The discrimination test developed for this procedure is common to several auditory aspects and includes loudness, narrowband speech compression, broadband music compression and spatial sound. For this discrimination task, different levels of a given perceptual aspect are illustrated by four stimuli carefully selected to ensure that two of them are clearly different while the two others are perceptually close to each other. In the headphone screening test, these sound samples were intended to illustrate four levels of spatial characteristics when replayed over stereo headphones.

Four dry monophonic speech samples with a different talker were selected and processed spatially to obtain the four levels required for this experiment. The two per-

ceptually distant stimuli were chosen to represent extreme cases of spatial sound distribution. For the lower level, a monophonic mix (*Mono*) of the four dry monophonic samples presented diotically was employed and for the higher level the Binaural Room Impulse Response (*BRIR*) processing approach was applied to assign each talker to a different position around the listener's head. For the two stimuli intended to be perceptually close in this GLS experiment, a stereo amplitude panning technique (*Stereo*) and a simplified Head-Related Transfer Function model (*HRTF*) were selected in order to obtain two intermediate levels of spatial separation, while preserving the anechoic nature of the original speech samples. These four spatial configurations are illustrated in Figure 5.6 with an idealized representation of the expected listening experience.

The four stimuli (referred to as 'systems' later in this section) were evaluated by 10 assessors from the sensory panel who developed the headphone sound consensus vocabulary. The test was run in a short session of 30 to 45 minutes organized in individual listening booths. The audio stimuli were loudness aligned beforehand and replayed at a level of 25 sones over the HD580 Sennheiser headphones. The *all samples - one attribute* presentation method described above was employed for the test administration and the user interface implemented on the GuineaPig2 listening test system included the audio exemplars anchored with an intensity derived from the results presented in the previous section. An illustration of the test user interface is shown in Figure 5.7 for the attributes *Tone color* and *Richness*. For each assessor, an independent random permutation was applied for the attribute presentation order and for the association between the letters A to D and the four audio samples.

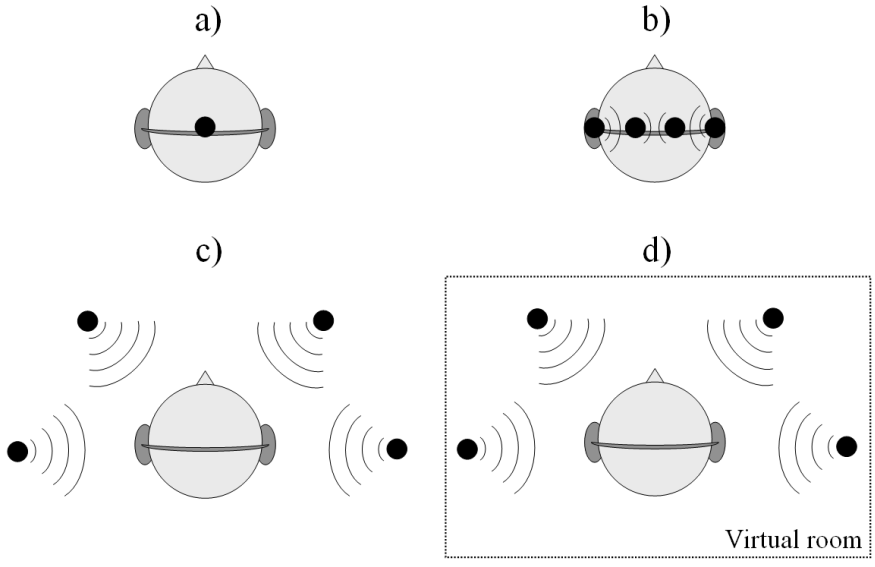
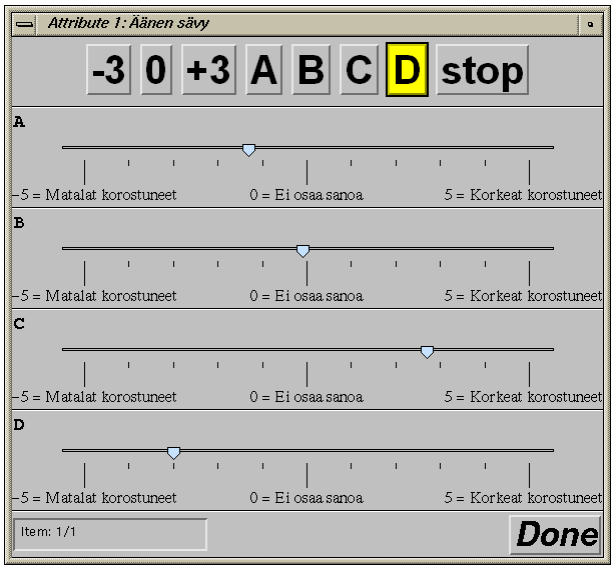
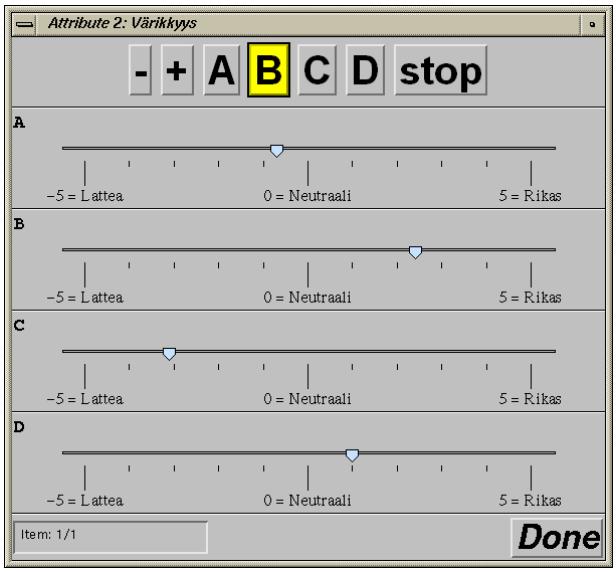


Figure 5.6: Idealized representation of the perceived location of four talkers when using the following headphone sound reproduction techniques: a) diotic replay (*Mono*), b) stereo amplitude panning (*Stereo*), c) simple head-related transfer function modeling (*HRTF*) and d) binaural-room impulse response processing (*BRIR*).



(a) Attribute *Tone Color*



(b) Attribute *Richness*

Figure 5.7: Two examples of GuineaPig 2 graphical user interface employed for the multiple stimulus presentation in this attribute rating test. The intensity anchoring used for the audio exemplars was derived from the listening test results presented in the previous section.

5.4.2 Result

The data resulting from a sensory profiling experiment can be analyzed in many ways including different statistical descriptions and visual representations of the attribute scores from either a univariate or a multivariate perspective. In the current example, two complementary approaches were selected to summarize the perceptual differences between the four stimuli under study, that is, a univariate visualization approach

referred to as a 'spider web' plot (Figure 5.8) and a principal component analysis (Figure 5.9).

Univariate view

Different types of graphical representations can be used to visualize the scores of a sensory experiment at an attribute level. Approaches include plots of mean scores and 95% confidence interval for each attribute separately or either 'Spider web' plots or 'Semantic Differential' Charts (Osgood, 1967) of several attributes simultaneously. The univariate statistical analysis supporting the visualization of differences between systems is usually based on the ANOVA framework.

Spider web plots are convenient to summarize the large amount of data resulting from attribute rating tests especially when the number of evaluated stimuli is small as in the present study. In this type of plots, each stimulus is visualized by a spider line on a number of spikes representing the different attributes of the sensory profile. The spider web plot presented in Figure 5.8 shows the mean attribute scores derived from the raw data of this sensory experiment. The attribute names are given with an indicator of the level of discrimination between the systems at a panel level, which was measured as a p-value for the F-ratio of the factor `System` resulting from a two-way ANOVA using the factors `System` and `Assessor`. It appears from this plot that only two attributes, *Balance of Sounds* and *Quality of Echo*, are not significant. It should also be mentioned that the attribute *Sense of Movement* was not included in this analysis because it has not been evaluated by all the assessors, some of them considering it irrelevant for this stimulus set.

Figure 5.8 illustrates clearly the overall similarity between the two systems *Stereo* and *HTRF*[8] and the larger differences between the two systems *Mono* and *BRIR*. The attribute *Broadness* shows the largest absolute difference between systems while the attribute *Sense of Distance* shows the most uniform distribution of mean scores. The perceptual differences between these four stimuli can be summarized in three main patterns: 1) attributes for which the system *Mono* differs from the other systems, e.g. *Sense of Direction* and *Ratio of Localizability*, 2) attributes for which the system *BRIR* differs from the other systems, e.g. *Tone Color* and *Richness*, and 3) attributes discriminating the systems *Stereo* and *HRTF* from the system *Mono* on one side and the system *BRIR* on the other side. This latter pattern is clear for example on the three attributes *Broadness*, *Balance of Space* and *Sense of Space*, which all relate to the perception of space.

Multivariate view

A principal component analysis (PCA) was also performed on the mean attribute data excluding the two non-discriminative attributes discussed above and the attribute *Sense of Movement*. PCA is a useful tool to explore the systematic variation of a multivariate data set and to highlight relationships between the objects and variables, that is, the systems and the attributes in our case. The idea of PCA is to find a set of linear combinations of the original (manifest) variables accounting for the maximum of

[8]Differences between the two systems *Stereo* and *HRTF* are significant for none of the attributes evaluated in this study.

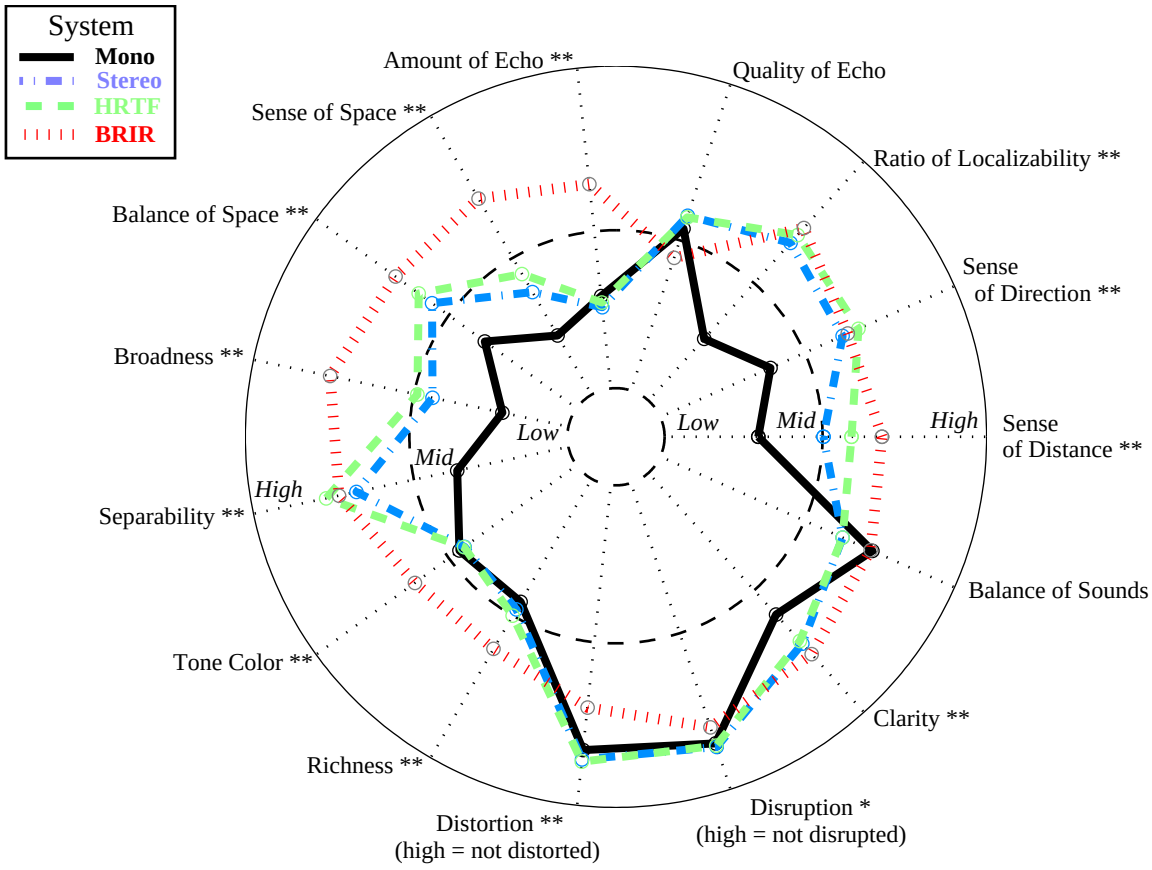


Figure 5.8: Univariate view of the mean sensory data with a 'spider web' plot. Significant levels for the *system* F-ratio are indicated as * for 0.05 and ** for 0.01. Low, Mid and High refer respectively to -5, 0 and +5 for the attributes *Tone color* and *Richness* and to 0, 5 and 10 for the other attributes.

variation in the data. The first few of these new (latent) variables explaining the most important part of the total variance are selected and the multivariate data exploration is made from the display of the objects and original variables in this low-dimensional subspace through their principal component scores and loadings respectively.

The interpretation of PCA results is often a somehow arbitrary task in practice and it can take different forms depending on where the focus is placed, i.e., either on the manifest variables or on the latent variables. The technique of predictive PCA biplot proposed by Gower and Hand (1996) was adopted for the small data set considered in the present study. This type of biplot aims at describing the objects directly in terms of the manifest variables using a technique of orthogonal projection on biplot axes with calibrated markers, which in our case correspond to the sensory descriptors with their original intensity scales. This technique illustrated in Figure 5.9 offers the closest multivariate equivalent of the univariate plot presented in Figure 5.8. An alternative approach often encountered in sensory data analysis aims at describing the latent variables in terms of the manifest variables, which is a process known as reification. Techniques to rotate principal components are often employed to simplify the loading structure and the subsequent interpretation as presented e.g. by Jolliffe (2002). In

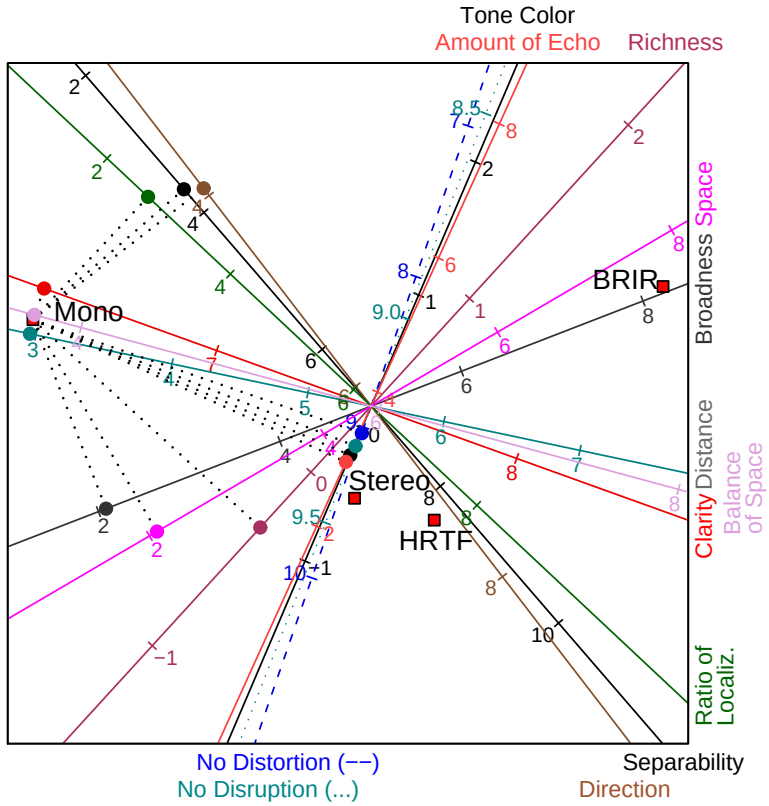


Figure 5.9: Multivariate view of the mean data with a predictive PCA biplot explaining 99.6% of variation in the data (82.2% for PC1 and 17.4% for PC2).

ideal cases, this approach allows to assign a sensory meaning to each component and describe the differences between objects in terms of these latent sensory dimensions.

The predictive PCA biplot presented in Figure 5.9 was created in the R statistical software with the BiplotGUI package (2008) designed to construct and interact with biplots of the type advocated by Gower and Hand (1996). The two first components of the PCA applied to the centered variables explained 99.6% of the variation in the data with 82.2% and 17.4% of explained variance for PC #1 and PC #2 respectively. The PCA scores define the relative position of the objects in the sub-space defined by the two first principal components. In the present case, the systems *BRIR* and *Mono* appear to lie further apart than the systems *Stereo* and *HRTF*. The original attributes are represented in Figure 5.9 by a set of calibrated biplot axes that brings additional information for the display interpretation. Firstly, these axes give a means to evaluate the co-variation between attributes, e.g. the attributes *Clarity*, *Distance* and *Balance of Space* form a group with an horizontal orientation on the display and the attributes *Direction*, *Separability* and *Ratio of Localizability* form another group with a diagonal orientation. Secondly, the technique of marker calibration employed in this type of biplot offers an intuitive way to assess the sensitivity of each attribute. This can be illustrated in the present figure by comparing the markers of the attribute *Broadness*, which span most of the intensity scale (from 2 to 8) on the display while the extent of

the attribute *Clarity* is only limited to the markers 7 and 8. The smaller range of the latter attribute highlights its lower contribution to the sample representation in this low-dimensional space. Thirdly, each system can be projected orthogonally on each axis on this biplot as illustrated for the system *Mono* in Figure 5.9. By exploiting this principle of projection, it was possible to identify the attributes discriminating best the four samples, that is, *Broadness*, *Sense of Space*, *Distance*, and *Balance of Space* in the present case.

5.4.3 Conclusion of this small sensory profiling experiment

The result of the small sensory profiling experiment reported above confirms the suitability of these four sound stimuli for the headphone spatial discrimination task of the GLS screening experiment. Indeed, the perceptual differences between these audio samples appear to relate principally to spatial aspects with a large contribution of the attributes *Broadness* and *Sense of space*. In addition, the distance between the four stimuli along this perceptual direction turns out to follow accurately the intent of the GLS experiment with two closely-spaced stimuli and two more distant stimuli. However, other differences in perceptual characteristics were also revealed from this sensory profiling test. In particular the departure of one stimulus from the three others was visible for the system *Mono* with respect to the attribute *Separability* and for the system *BRIR* with respect to the attribute *Amount of echo*.

5.5 Discussion

The implementation and outcome of the consensus vocabulary development work presented in this chapter are discussed below with an explicit attention to the sound stimulus selection, the vocabulary development procedure, the vocabulary generated by the panel, the selection of sound exemplars to illustrate and anchor the attributes, the attribute rating methodology and finally the illustrative application of this vocabulary.

The selection of audio stimuli for this descriptive analysis study aimed at covering the main headphone listening scenarios encountered in practice with a limitation to static rendering conditions. The classification presented in Table 5.1 provided a useful basis for this sample selection. The stimulus selection procedure employed in this work differs significantly from the approach found in the study of loudspeaker sound reproduction systems by Zacharov and Koivuniemi (2001c) in which a set of factors defining systematically the sound source, the acoustic space and the recording technique was employed. The present selection process resulted in a less uniform set of stimuli than in the study of Zacharov and Koivuniemi (2001c) but was found to be appropriate for the intended descriptive analysis task. Also, the careful perceptual screening applied to the large number of initial audio samples ensured that a comprehensive set of stimuli covering the full sensory space was chosen. It should be noted however that this procedure could have been improved by the use of unbiased listeners for the sample screening process for example with the help of a sound stimulus grouping task.

The methodology employed for the vocabulary development of this study followed the procedure developed by Zacharov and Koivuniemi (2001b) and Mattila (2001a) but included several modifications intended to optimize the process, that is, a shorter initial elicitation work, a step of panel discussions in small groups and a final phase combining individual usage of the attributes and panel discussion to refine the consensus between the assessors. The overall process was reduced to a total duration of 20 hours per assessor (plus the panel leader time) but the author is confident that further optimization of the procedure could be achieved. For example, a faster individual elicitation step of 2 hours instead of 4 hours might be sufficient while a larger effort could be spent on the vocabulary alignment using a more systematic procedure for providing feedback to the assessors on their attribute usage performance. Additional time reduction might also come from the added sensory expertise of assessors, the panel leader know-how and the automation of the tools supporting the consensus alignment work.

The vocabulary agreed on by the panel during this verbal elicitation experiment comprises sixteen sensory descriptors organized in three perceptual groups. The diversity of these descriptors reflects the complexity of the stimulus set selected for the study and illustrates the multivariate character of spatial sound reproduced over headphones. A wide range of perceptual aspects is covered with an important spatial component relating to the localization and externalization of sound and the description of space but also with other components relating to timbre, loudness and artifacts. The combination of spatial and non-spatial aspects seen in this vocabulary is in line with the results from of previous works on spatial sound perception reviewed in Chapter 3 (see Table 3.1).

Despite the different nature of the sound reproduction scenarios found in the studies reviewed in Section 3.3.3, clear similarities in perceptual description can be observed. The distinction between source-related and room-related attributes noted earlier applies to the current study and it also appears that most of the groups defined loosely from the literature review are represented in this headphone vocabulary, e.g. *tone color* for the timbral group, *Separability* for the timbral discrimination group, *Balance of sounds* for the loudness group, etc.. It can be noted however that in the present case sound sources are not described individually but at a more general level with the attribute *Ratio of localizability*.

The relationship of this headphone vocabulary to conventional descriptive systems applied in perceptual studies on headphone sound is also apparent to some extent although not in a direct manner. The sound localization aspect is covered at a more abstract level than the coordinate system approach commonly found in sound localization studies to quantify e.g. azimuth and elevation. It should also be noted that the sensory descriptors of the localization group describe a 'sense' rather than an 'amount' of *Distance* and *Direction*. The present vocabulary covers the important issue of externalization in headphone sound with the attribute *Broadness* but does not include the aspects of front-back confusion and (exaggerated) sound image elevation commonly encountered in headphone studies. Interestingly, the panel identified the issue of front-back ambiguity during the vocabulary development but did not agree on its description because assessors experienced this phenomenon differently. This is also

partly the reason why localization attributes such as *Distance* or *Direction* were defined by the panel through the concept of 'definability'. A similar type of description has been adopted for example by Guastavino and Katz (2004) in their study of sound reproduction systems with multiple spatial dimensions for the attribute *Readability* defined as 'the spatial definition, readability of the scene' and with the verbal anchors 'poorly defined' and 'well defined'.

The literature review on CV methods provided in Chapter 4 highlighted the importance of physical references in a vocabulary development work. Such attribute anchors can help assessors to get a common understanding of the attributes employed to describe perception and they contribute to aligning the way they use the attribute scales. In the present study, an effort was made at the end of the attribute development to select appropriate sound exemplars providing intensity anchors for each descriptor. Although this task was useful to some degree in the attribute rating phase, the resulting quantitative anchoring system was not entirely satisfactory because the panel was unable to agree on the intensity anchors of six attributes (Table 5.2). This result indicates that more effort should have been put on this aspect during the final stage of the vocabulary development procedure. It should also be emphasized that sound exemplars are important in this type of experiment not only for the sensory assessors developing or learning the vocabulary but also for the purpose of communicating the meaning of the attributes to outsiders.

The investigation on attribute rating methods reported in this chapter highlighted an important distinction between the absolute nature and relative nature of the cognitive task associated with a single and a multiple stimulus evaluation respectively. The 'multiple stimulus - single attribute' presentation method was selected in this study for its superior sample discrimination ability as demonstrated in the small comparative test reported above. This choice is supported by the findings of Ishii *et al.* (2008) that comparative grading is better suited for assessors with a moderate level of training. This result implies however that more effort would have been required with the headphone panel to achieve a good discrimination when using a single stimulus presentation method. Regarding the inclusion of sound exemplars in the comparative attribute rating windows of the sensory profiling experiment reported in Section 5.4, the author speculated that this approach allows to preserve the absolute character of the stimulus evaluation but this would deserve a formal verification. It should also be mentioned that the difficulty to evaluate simultaneously a large number of stimuli, e.g. more than 15, represents a serious limitation of this comparative rating approach.

Finally, the small sensory profiling experiment performed on the four GLS headphone sound samples illustrated the overall applicability of the headphone vocabulary. It can be expected that developing a vocabulary specifically for these four stimuli would give a very different set of sensory descriptors but applying a multivariate analysis such as PCA to the resulting attribute data would most probably lead to a sensory map quite similar to the one obtained in the present experiment. The large co-variation between many attributes seen in Figure 5.9 is mainly due to the relatively large size of the vocabulary and the limited perceptual dimensionality of this small

data set. However, it appears that the generic vocabulary for headphone sound reproduction allowed to unravel the two or three essential perceptual aspects describing this stimulus set.

Chapter 6

Application of the consensus vocabulary to spatial enhancement systems

6.1 Introduction

Following the phase of consensus vocabulary development described in the previous chapter, the resulting set of attributes was applied to a specific headphone application referred to as 'spatial enhancement'. The aim of these signal processing techniques is to enhance the spatial characteristics of audio material reproduced over headphones. The present chapter starts with a brief overview of the audio stimuli selected for this attribute rating test and a description of the experimental procedure. The data resulting from this experiment is then analysed in details with both a univariate and a multivariate approach.

6.2 Presentation of the experiment

6.2.1 Spatial enhancement algorithms and music clips

The set of audio stimuli selected for this experiment relates to the spatial enhancement systems for headphone reproduction evaluated with a set of preference tests briefly reported in Chapter 3 (Section 3.4). Two groups of algorithms were chosen including eight spatial enhancement algorithms for music reproduction over headphones and eight virtual 5.1 algorithms for headphone reproduction. However, the focus of this chapter is placed on the analysis of the CV experiment relating to the first set of algorithms.

The set of eight reproduction techniques considered in the present study, which are referred to as 'systems', includes the stereo unprocessed material (1_{S}), a monophonic down-mixed version of the stereo material (2_{M}), a spatial enhancement algorithm for monophonic music material (4_{Me}) and five stereo enhancement algorithms representing different perceptual flavors comprising an algorithm based on the virtual room processing approach (3_{Se}), an HRTF-based algorithm (6_{Se}) and three algorithms em-

ploying alternative processing methods (5_{Se}, 7_{Se}, and 8_{Se}). The selection of these reproduction techniques aimed to cover a broad range of perceptual aspects including both spatial and timbral differences. A set of music excerpts by *Scritti Politti*, *Madonna* and *Tuck & Patti* was also selected and processed with these eight systems. These three program items will be referred to as the 'music clipa' #1, #2 and #3 respectively. This selection gave a total of 24 sound stimuli organized in '8 systems $\times$ 3 music clips' for the attribute rating test reported in the present chapter. For more details about the spatial enhancement algorithms and music clips, the reader is referred to Lorho (2005a).

6.2.2 Listening test administration

Chronologically, the experiment started with the sensory evaluation of the music clip #2 and a month elapsed before the study could be continued due to practical constraints. Therefore, it was decided to organize a short test session to refresh the vocabulary to the listening panel with a small set of stimuli before continuing with the main study of the spatial enhancement systems. Note that the audio sample selection for this attribute refresher task and the resulting sensory profile have been presented in Chapter 5 (Section 5.4). Then, the main listening test was run over a period of about a month and included the 6 following series of stimuli: the music clip #3, two program items relating to the experiment on virtual 5.1 algorithms for headphones and the music clip #1 repeated 3 times. The order of presentation of the different items was randomly permuted with the constraint that no repetition of the music clip #1 would happen consecutively.

The listening test administration of this study was identical to the procedure reported in Chapter 5. A multiple stimulus presentation method including eight systems listed from A to H was employed, as illustrated by the GuineaPig2 user interface window shown in Figure 6.1. A random permutation was applied for the letter association of each series of eight stimuli and for the attribute presentation order. The interface allowed for direct switching between the test stimuli using a 20ms exponential crossfade with a switching latency of 200ms. The user interface window of each attribute also included the two (or three) associated attribute exemplars selected during the consensus vocabulary development experiment to provide the subjects with an absolute reference. For example, in Figure 6.1, the attribute *Clarity* includes two sound exemplars with the numerical anchors 1 and 9. Each series of eight stimuli was evaluated in one test session which lasted around 60 to 90 minutes and 10 assessors from the consensus panel of this headphone study participated to the experiment.

6.2.3 Overview of the resulting set of data

The outcome of any quantitative attribute rating test performed with a sensory panel represents a large amount of data. The set of scores collected in a single consensus vocabulary profiling test made on I objects with J attributes by K assessors can be represented as a three-dimensional array comprising three modes of classification as illustrated on the left side of Figure 6.2. This format is applicable to the attribute

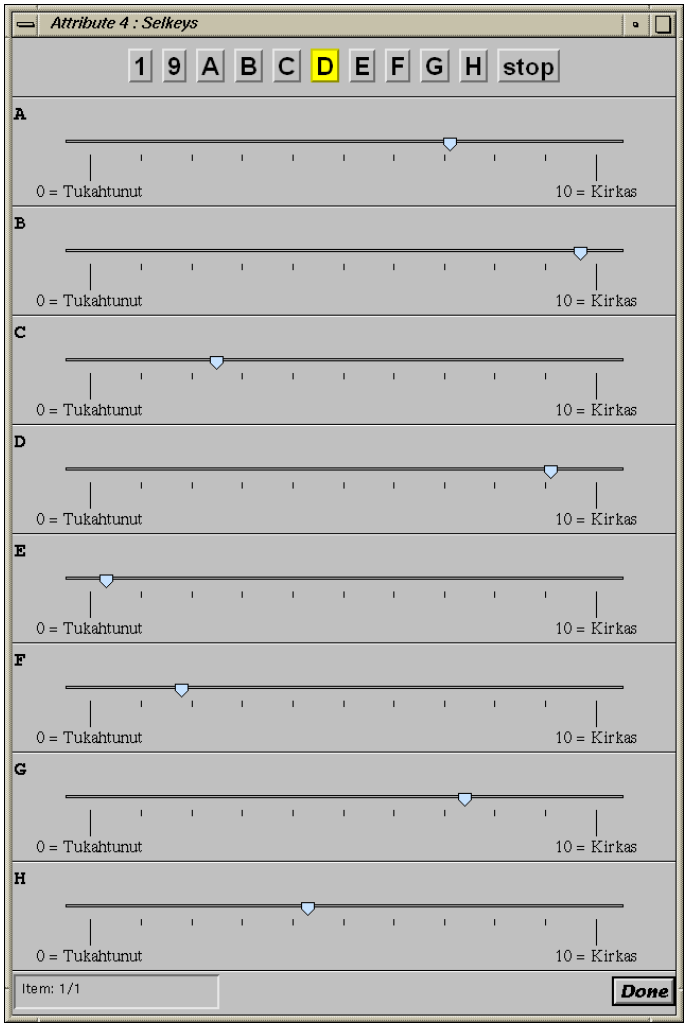


Figure 6.1: Example of GuineaPig 2 user interface employed for the attribute rating test.

scores resulting from a CV experiment because the variables of the mode B are considered to be commensurable, that is, the attribute scales are assumed to have a similar meaning to all the assessors of the panel. The term 'slab' is employed in this thesis to refer to a two-way sub-array of this data structure. For example, a 'frontal slab' represents the sensory profile of one assessor as illustrated on the right side of Figure 6.2. Following the same principle, a 'lateral slab' would contain the scores of all the assessors for one attribute.

The data relating the present experiment contains several sets of sensory profiles. The upper part of Figure 6.3 illustrates the structure of the full set of raw scores. The outcome of each series of music clip evaluation made by the panel can be represented by the three-way array defined earlier. Each frontal slab contains the sensory profile of an assessor, that is, a matrix of 8 *systems* by 16 *attributes*. As the panel repeated three times the evaluation of the music clip #1, the associated data set of 3840 data points might be represented as a four-way array of size 8 algorithms $\times$ 16 attributes

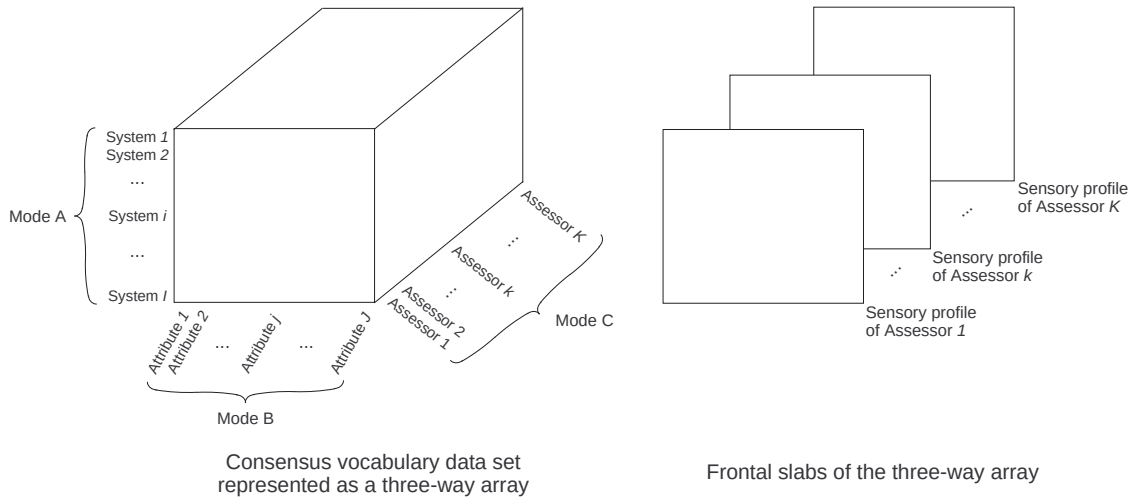


Figure 6.2: Representation of the sensory data set as a three-way data array.

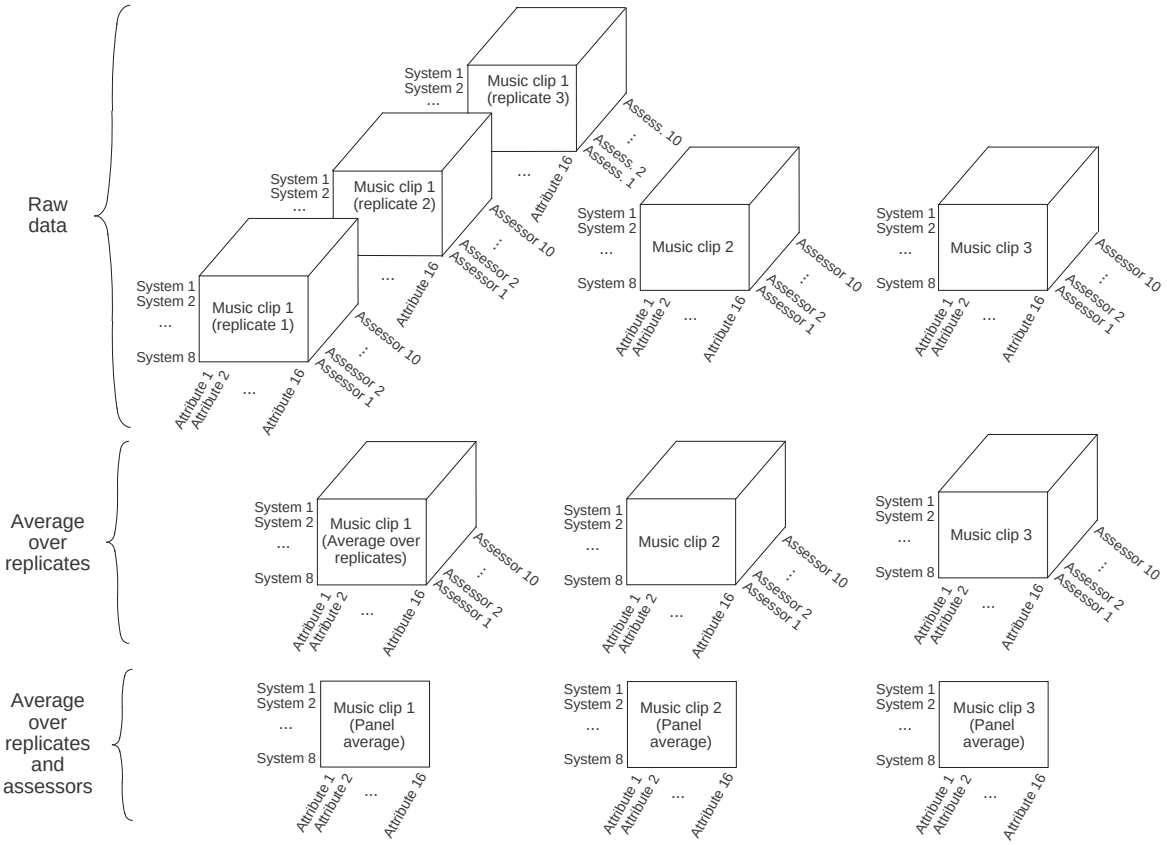


Figure 6.3: Illustration of the full set of attribute rating scores resulting from the study of spatial enhancement systems.

× 10 assessors × 3 repetitions but for visualization purpose, this data structure is illustrated by three different three-way arrays on the left of the top row in Figure 6.3. Each of the two other music clips is represented by a single three-way array because they have not been replicated in this attribute test.

Averaging over the triplicates of the music clip #1 is straightforward and leads a three-way array, which reduces the representation of the full experiment to three comparable arrays as shown in the middle row of Figure 6.3. Additionally, by averaging each of these three-way arrays across the (levels of the) mode `Assessor` after possibly some pre-processing on the three-way arrays, a two-way matrix of size 8 *systems* by 16 *attributes* is obtained for each music clip as shown in the lower row of the figure. This averaging process assumes that attributes measure the same sensory characteristic across assessors, which is justified in a CV experiment.

The analysis of this type of repeated attribute rating data can have different aims. Information can be collected either on the measured entities, that is, the systems characterized by the sensory descriptors, or on the measurement tool, that is, the consensus vocabulary and its associated panel. The investigation reported in this chapter focuses on the former perspective and for this reason an average over the triplicates of the music clip #1 was applied to obtain a three-way array similar to the two other music clips. A study of these three data sets was then considered with different types of analysis techniques.

6.3 Analysis of results

The three arrays shown on the middle row of Figure 6.3 were used as a starting point for analyzing the result of this perceptual evaluation of stereo enhancement algorithms for music reproduction over headphones. First, each music clip was considered separately and analyzed by a univariate and a multivariate approach and then an analysis was performed on the three music clips combined.

6.3.1 Analysis per music clip – Univariate approach

The univariate approach employed for the separate music clip analysis is based on an ANOVA applied to each attribute using the factors `Assessor` and `System`. Plots of estimated means and 95% confidence intervals were also derived from these ANOVA results for the factor `System` and were studied for each attribute.

Analysis of variance

The analysis of variance framework is frequently used for the univariate statistical analysis of sensory profile data (Lea *et al.*, 1997). The term 'univariate' here means that only the scores collected for one attribute are considered at once, which corresponds to either a lateral slab of the three-way array in Figure 6.2 or one lateral slab per three-way array when replicates are present.

The ANOVA model can include the factor `Product` defining the entities under study, the factor `Assessor` and possibly the factor `Session` when the replicated at-

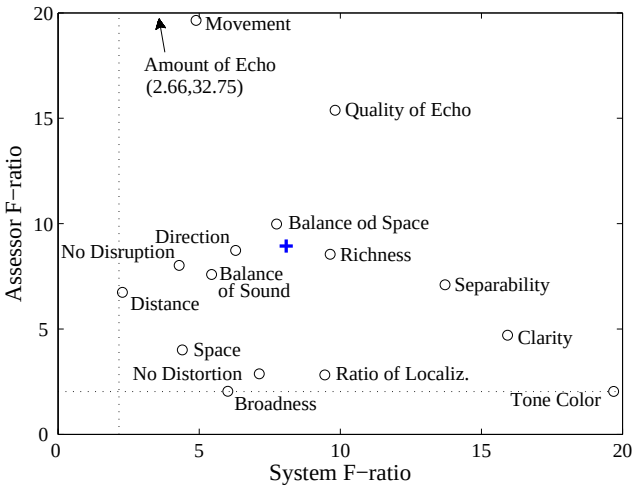
tribute ratings have been run at different occasions. The selection of terms to be included in an ANOVA defines how the data is being modeled and has an impact on the main output of the analysis, that is, the Fisher ratio (F-ratio). When repetitions are present, a two-way ANOVA model can be applied in which the `Product` term models the product differences, the `Assessor` term models the level differences in scoring between assessors and the `Product` $*$ `Assessor` interaction term models the differences in relative product scoring between assessors, while the residual noise models the individual variability. The F-ratio gives an indication of the significance level of the different terms but it should noted that selecting the factor `Assessor` as a fixed effect or as a random effect has some influence on the resulting `Product` F-ratio (Næs and Langsrud, 1998). Most often, a mixed ANOVA model is selected in which the factor `Assessor` is considered random while the other terms are considered fixed. This choice implies that the `Assessor` main effect and more importantly the `Product` main effect are tested against the `Product` $*$ `Assessor` interaction.

The model selection depends somewhat on the application and the question to be answered. From the perspective of the panel leader, the interest of the analysis is usually to assess the performance of the sensory panel and therefore an ANOVA model decomposing the source of data variation into all the components mentioned above would be the most appropriate as it would provide a detailed view of the assessor differences. Note that more advanced ANOVA models might also be adopted for this purpose as for example the model of Brockhoff (2003) described in Chapter 4. However, a different perspective could be to focus exclusively on the global product differences representing the most important experiment outcome. In that case, a simple one-way ANOVA model with the factor `Product` might be the most conservative choice as the residual would then include all the composite variability of the data provided by the sensory panel.

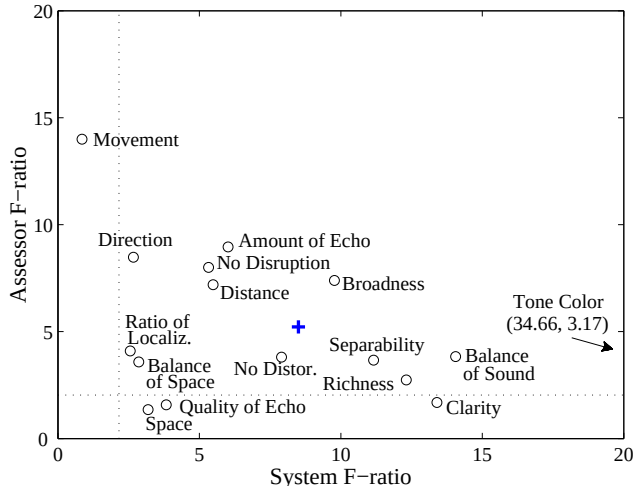
The analysis considered in this section is based on the set of data shown in the middle row of Figure 6.3 which does not include repetitions. An ANOVA was applied for each attribute and each music clip separately. Due to the absence of repetitions, no separation can be made in the ANOVA model between the assessor variability (the residual noise) and the individual differences (the `System` $*$ `Assessor` interaction). This means that the `Assessor` and `System` main effects are tested against a composite residual including both individual variability and assessor disagreement.

The F-ratio of the factors `Assessor` and `System` obtained for each attribute was plotted on a scatter plot for each music clip as illustrated in Figure 6.4. A cross representing the average F-ratio of the 16 attributes for each factor has been added for each music clip as well as an horizontal and vertical line illustrating the 5% F-ratio for these two factors. While the higher the `System` F-ratio the better in the present analysis, the opposite applies for the `Assessor` F-ratio because a significant level for this factor means that assessors vary in their level of scoring. The `Assessor` F-ratio gives therefore an indication of the validity of the absolute attribute anchoring implemented in this sensory experiment.

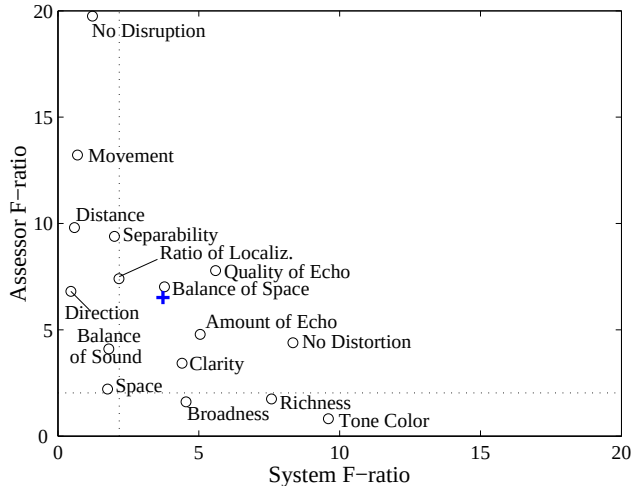
The three scatter plots presented in Figure 6.4 give a global view of the level of discrimination between systems for the different attributes and music clips. It can



(a) Scatter plot of F-ratios for the music clip #1



(b) Scatter plot of F-ratios for the music clip #2



(c) Scatter plot of F-ratios for the music clip #3

Figure 6.4: Scatter plots of `System` F-ratio versus `Assessor` F-ratio for the three music clips included in this study. The blue cross in the three graphs represents an average F-ratio over the 16 attributes.

be seen from these graphs that `System` F-ratios are significant for all attributes in the music clip #1 test, for all attributes except *Sense of movement* in the music clip #2 test and for only nine attributes in the music clip #3. The `Assessor` F-ratio of most attributes also appear to be significant in these three graphs, which means that assessors did not use the same level of scoring despite the absolute attribute anchoring technique described in Chapter 5. It can also be noted from these graphs that the attribute *Tone Color* has the largest F-ratio and a relatively low `Assessor` F-ratio for the three music clips. Comparing the F-ratio averaged over the 16 attributes for the different music clips (shown as a cross in the three scatter plots), the music clip #3 can be seen to have the lowest mean `System` F-ratio due to the large number of non-discriminant attributes while the music clip #1 has the largest average `Assessor` F-ratio partly due to the large F-ratio of the attribute *Amount of Echo*.

Plots of system differences per attribute

The ANOVA study presented above showed that the individual scoring level has a large influence on the attribute ratings in this experiment. Plotting the mean attribute scores of the different systems directly from the raw data as usually applied in descriptive statistics would be problematic in the present case because the large variability due to the level difference between assessors might mask the relative differences between systems at a panel level. For this reason, the author preferred plotting the estimated system mean scores and 95% confidence interval (CI) from the ANOVA models presented above, which does not include the source of variation due to the main effect `Assessor`.

Figure 6.5 presents the average attribute scores of the eight systems for the music clip #1 and similar plots for the two other music clips are shown in Appendix A (Figures A.1 and A.2). In these graphs, the error bars around the means are computed from the error term of the ANOVA model described above and they give an indication of the level of disagreement between the assessors for each attribute. Clear differences in error bar size can be observed between some of the attributes. For example, the panel appears to be more in agreement for the attribute *Tone Color* than for the attribute *Balance of Space* on the three music clips. The non-discriminant attributes are marked with the label 'n.s.' in these graphs and it can be noted that three of them show relatively large error bars, namely the attributes *Distance*, *Direction* and *Space* of the music clip #3. A comparison of the three music clips shows that the error bars of the music clip #1 tend to be slightly smaller than for the other clips overall. It should be recalled however that this set of scores represents an average over three replicates which might have reduced the individual part of the variability in scoring. It also appears from these series of graphs that a larger range of scales was employed for the second music clip, especially with the timbral attributes.

Focusing on the mean plots of the music clip #1 in Figure 6.5, important differences are visible between the stereophonic-based systems and the monophonic-based systems. For example, the unprocessed stereo (1_S) and the monophonic down-mixed version of the stereo material (2_M) appear to be clearly discriminated by the panel both with the group of localization attributes, i.e., *(Sense of) distance*, *(Sense of)*

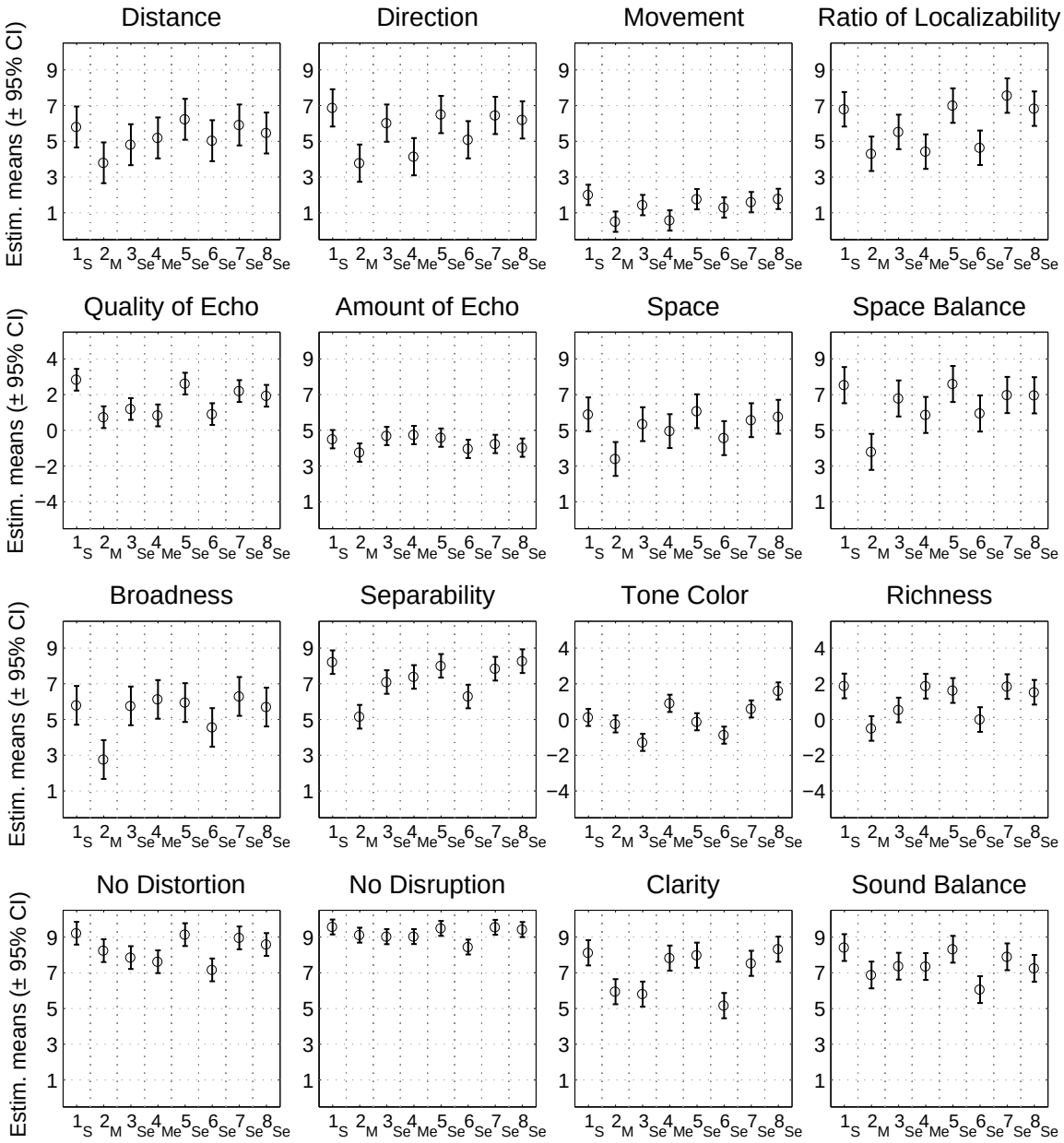


Figure 6.5: Mean scores estimated from ANOVA and 95% CI of the eight spatial enhancement systems on 16 attributes for the music clip #1.

direction, *(Sense of) movement* and *Ratio of localizability* and the group of spatial attributes, i.e., *Quality of echo*, *(Sense of) space*, *Balance of space*, *Broadness* and *Separability*. Also, the mono-to-3D algorithm (4_{Me}) is seen to resemble more the monophonic reproduction (2_{M}) in terms of localization attributes and more the stereo reproduction (1_{S}) in terms of spatial attributes.

The mean plots of the music clip #2 (Figure A.1 in appendix) show a similar trend between the three systems 1_{S}, 2_{M} and 4_{Me} but this is less systematic within attribute groups. Additionally, it appears that the discrimination pattern between the eight systems is more complex with this music clip for the group of timbral attributes, i.e., *Tone color*, *Richness*, *Distortion*, *Disruption*, *Clarity* and *Balance of sound*.

The results of the music clip #3 (Figure A.2 in appendix) are only significant for nine attributes. Clear differences can be seen between some of the systems for most of the timbral and spatial attributes but the discrimination level of this stimulus set is very low for the localization attributes. A significant difference can be seen between only two systems for one attribute in this group, that is 4_{Me} and 6_{Se} for *ratio of localizability*. Also, the discrimination between the mono- and stereo-based systems is less clear than for the two other music clips overall. It should be noted however that this musical excerpt is very different from the two others. While the two pop music tracks include clear stereo panning effects, this jazz tune is a subtle instrumental recording with a slightly narrow soundscape, which can justify the lower level of perceptual differences observed between systems for the localization attributes.

It can be noted that the mean plots presented in these three figures give a concise overview of the differences between systems at an attribute level but they are not very effective for visualizing the global sensory profile of each system or for comparing two or more systems on several attributes. The spider web plot approach presented in the previous chapter or more generally the sensory profile plot would be better suited to visualize the full set of attribute scores of a given system and to compare the sensory profile of different systems. However, the readability of such plots can also be significantly reduced when a large number of systems is displayed. Interactive visualization tools would definitely be beneficial for this type of sensory profiling data visualization.

In summary, the learning from this analysis was that the range of scores employed at a panel level does not cover the full intensity scale for most attributes while individual differences in scoring remain large despite the correction of assessor level differences. This might be a consequence of the limited panel training but it can also be questioned whether the attribute anchoring approach employed for the present experiment was optimal or not. Indeed, the set of sound exemplars selected for the generic vocabulary of headphone sound reproduction covers a wider perceptual range than the stereo enhancement systems tested in the present study.

An additional issue not mentioned earlier but visible to some extent in the mean plots concerns the relatively large degree of co-variation between some of the attributes for the three music clips, especially for attributes of the same group. This aspect can be investigated through an analysis of correlation between attributes, as presented by Lorho (2005a) for the same data set, but it will also become clear from the multivariate data analysis presented next.

6.3.2 Analysis per music clip – Multivariate approach

The multivariate exploration of sensory profiling data is commonly performed via a principal component analysis (PCA) applied to the two-way matrix of I objects by J attributes containing panel mean scores. Individual sensory profiles can also be exploited within the PCA framework to add information about the panel reliability as discussed in Chapter 4 (Section 4.4). The use of a three-way pre-processing technique are described and justified in the first part of this section and the result of a PCA applied to the average sensory profile of each music clip is then reported.

Data pre-processing

Several issues with the sensory profiling data set considered in this study were highlighted in the previous section. This includes the level differences between assessors seen from the ANOVA study and the scaling differences between attributes at a panel level seen from the mean score plots. In addition, the risk of scaling differences between assessors should be investigated. A pre-treatment of the raw attribute scores can be applied to eliminate or at least reduce these individual variations specific to scale usage while preserving the relative differences between systems. This procedure has also the purpose of improving the quality of multivariate model subsequently applied.

Pre-processing includes the two steps of centering and scaling. These two data treatments are most effective when applied directly the original individual sensory data and they can be defined formally from the three-way array representation and terminology shown in Figure 6.2. Centering in a three-way array is usually applied 'across' a given mode. For example, centering across the mode `System` is computed as

$$\tilde{x}_{ijk} = x_{ijk} - \bar{x}_{.jk} \tag{6.1}$$

where the dot subscript is employed to indicate the mean across $i = 1, \ldots, I$.

Scaling (or normalization) can be applied in several ways. The most common approach for a three-way array is to scale 'within' a given mode. For example, scaling within the mode `Attribute` is achieved by computing the scaling factor

$$\nu_j = \sqrt{\sum_{i=1}^{I} \sum_{k=1}^{K} x_{ijk}^2} \tag{6.2}$$

and computing the normalized data as

$$\tilde{x}_{ijk} = \frac{x_{ijk}}{\nu_j} \tag{6.3}$$

This scaling method is preferred in the context of three-way data modeling because it has the property of preserving the structure of the slab within which the scaling takes place (Bro and Smilde, 2003). This is not the case when considering the alternative scaling approach applied 'across' a given mode. For example, scaling across the mode `System` would be computed as

$$\tilde{x}_{ijk} = \frac{x_{ijk}}{\sqrt{\sum_{i=1}^{I} x_{ijk}^2}} \tag{6.4}$$

The latter approach is similar to the two-way array scaling commonly applied in sensory studies. Also, when combining the two steps of centering and scaling across the mode `System`, we obtain the usual 'z-score' normalization. It should be noted however that applying a z-score normalization directly to the two-way matrix of the panel mean sensory profile does obviously not correct the individual differences but it affects the data by giving a similar weight to all attributes, including those for which system means are not significantly different.

Different types of data pre-treatments can be applied and combined in practice. As the selection of a specific procedure depends on the intended effect, it is therefore very important to fully understand the impact of the pre-processing step(s) on both the attribute scores and the multivariate model output.

The aim of pre-treatment in the present case was to correct the issues mentioned above. The first aspect of assessor level effect was handled by centering the three-way array across the mode `System`, which is an approach commonly used in sensory studies.

The second aspect of attribute scaling differences was handled by scaling the three-way array within the mode `Attribute`. This type of normalization affects equally all individual sets of scores for a given attributes but it destroys the absolute scaling of the different attributes, which is acceptable in the present study. The role of this scaling treatment is only to balance the contribution of each (discriminant) attribute in the PCA.

For the third aspect of individual scaling differences, several approaches were considered in this study. A conventional approach would consist in scaling the three-way array across the mode `System`[1]. However, a recent publication by Romano *et al.* (2008) illustrated that this simple standardization method is less effective in reducing individual scaling differences than the assessor modeling method Brockhoff and Skovgaard (1994); Brockhoff (1998) based on a complex ANOVA model and the 'ten Berge scaling' method (ten Berge, 1977) based on minimizing the difference between each combinations of two assessors. The ten Berge scaling method was applied to the three music clips separately and it was found to improve the overall discrimination level in all cases. With the music clip #1 for example, applying the ANOVA study described above to the pre-processed data gave a mean F-ratio over all attributes of 8.0, 9.9 and 11.7 for the raw data, the standardized data and the data scaled with the ten Berge scaling method respectively. This technique was therefore selected to address the third issue of individual scaling differences in this study.

Following this multi-stage three-way array pre-treatment, a new univariate ANOVA was run on each attribute to identify and discard attributes for which system means were not significantly different at a panel level. The result was similar to the one found in the previous ANOVA analysis with 16, 15 and 11 discriminant attributes for the music clips #1, #2 and #3 respectively. Finally, a two-way matrix of mean scores over the panel including only these attributes was created for each music clip.

[1]Note that if this scaling approach was applied, the second second problem of attribute scaling differences would be covered automatically.

Visualization of principal component analysis results

The two-way matrix of mean scores resulting from the data pre-treatment described above was submitted to a PCA for each music clip. This multivariate data analysis technique was introduced in Chapter 4 with a description of several approaches to visualize PCA results. A conventional visualization approach employing the score/loading plots was applied in the present analysis to explore the systematic structure of the data. The number of relevant components to select was judged from the relative increase in cumulative variance when adding new principal components[2] and the results were interpreted in this low-dimensional subspace. Note that the figures of the PCA plots relating to the music clip #1 are visible in this section while the figures of the two other music clips can be found in Appendix B.

In this analysis, the PCA score plots show the sensory map of the systems under study, that is, a display in which object positions reflect their inter-distance in the sub-space defined by the principal components. The ellipses included in these plot correspond to the 95% confidence regions around the system scores and represent the individual variability of the panel mean. They are computed from the pre-processed individual sensory profiles using the bootstrapping method proposed by Husson *et al.* (2005) using a procedure that can be summarized as follows.

- create a virtual sensory panel by bootstrapping of the original panel and compute the associated average sensory profile,
- project this virtual average sensory profile on the principal component space,
- repeat the two previous steps of this procedure a large number of times (e.g. 5000),
- for each system, draw an ellipse summarizing the bivariate distribution of virtual mean scores in the two-dimensional subspace of interest (PC#1 and PC#2 in the present case).

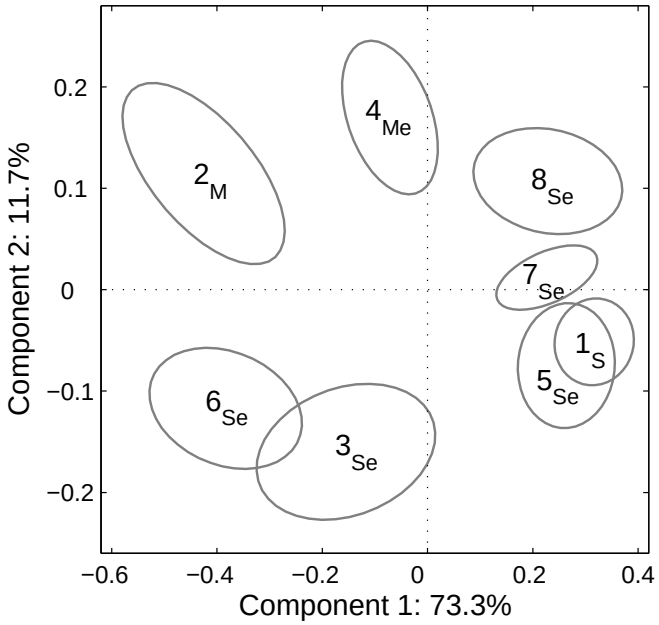
The PCA loading plots are complementary to the score plots and illustrate the relationships between the original variables (i.e., the attributes) and the principal components displayed. In this type of plots, the longer the vector, the more important the corresponding attribute is for discriminating the samples. For clarity, the attributes with the largest loading values have been highlighted by thicker lines in these plots.

PCA of the music clip #1

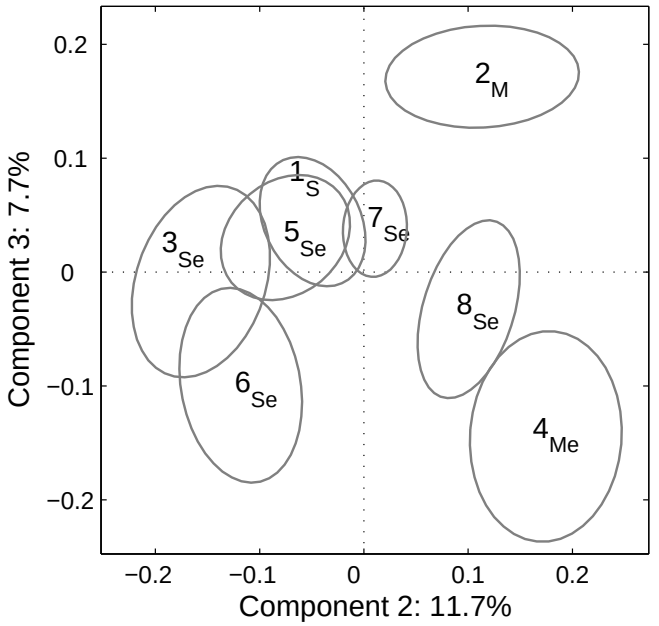
A 3-component model explaining 92.7% of the variation in the data was considered appropriate to describe the consensus data set of the music clip #1. These components explained respectively 73.3%, 11.7% and 7.7% of the total inertia while each of the remaining components accounted for less that 5%. Figures 6.6 and 6.7 show the score plots and loading plots of the PC 1–2 subspace and the PC 2–3 subspace.

The level of separation between the ellipses in the score plots (Figure 6.6) indicates that the panel was able to discriminate most of the systems. Some overlap can be seen

[2]Kasier's criterion stating that only eigenvalues superior to 1 should be retained is not applicable in this case because the variables have not been standardized directly from the two-way matrix.

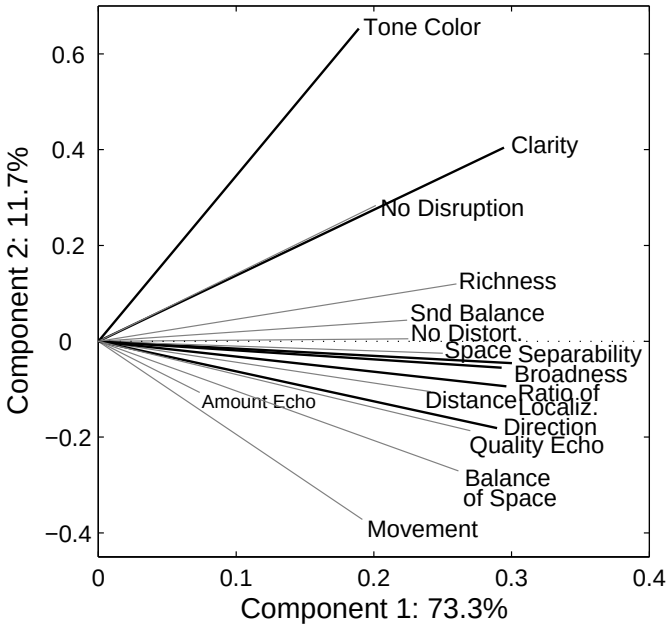


(a) PCA scores - PCs 1–2

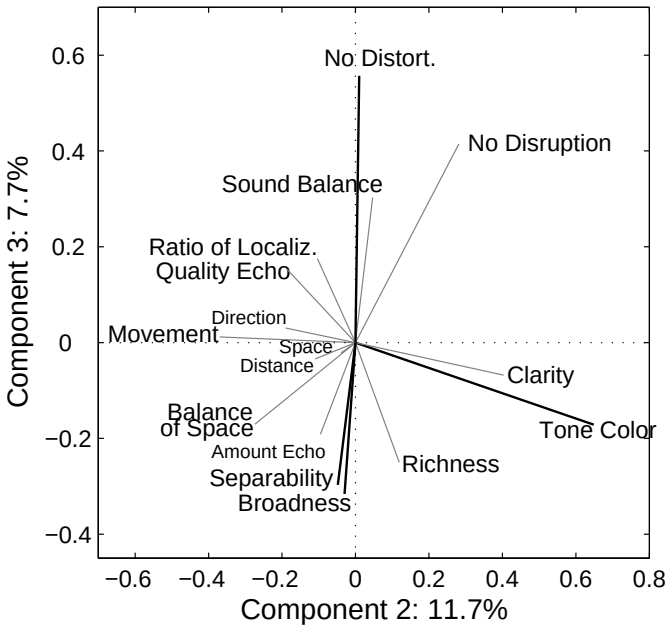


(b) PCA scores - PCs 2–3

Figure 6.6: Three first components of a PCA applied to the music clip #1. Ellipses around the PCA scores in the upper plot represent 95% confidence regions based on assessor variability.



(a) PCA loadings - PCs 1–2



(b) PCA loadings - PCs 2–3

Figure 6.7: Two first components of a PCA applied to the music clip #1. These loading plots illustrate the relationship between the attributes and the PC dimensions.

however between the unprocessed stereo (1_S) and the stereo enhancement algorithm 5_{Se} and also between the two stereo enhancement algorithms 6_{Se} and 3_{Se}. The loading plots (Figure 6.7) illustrates the large co-variation of the attributes with a first principal component explaining the most important part of the variance while the other dimensions are loaded by only few attributes like *Tone Color* for PC #2 and *(no) Distortion* for PC #3.

This sensory map can be interpreted broadly from these two sets of plots as follows. Firstly, the monophonic reproduction (2_M) and the stereo enhancement system 6_{Se} are clearly opposed to the unprocessed stereo (1_S), the systems 5_{Se}, 7_{Se} and 8_{Se} in terms of the localization attributes *Direction* and *Ratio of Localizability*, the spatial attribute *Broadness* and the timbral attribute *Separability* mainly but also other attributes loaded to a smaller extent on PC#1.

Secondly, a perceptual direction defined by the timbral attributes *Tone Color* and *Clarity* in the upper part of Figure 6.7(a) and in the right part of Figure 6.7(b) separates the mono-enhancement system (4_{Me}) and the system 8_{Se} from the stereo enhancement systems 3_{Se} and 6_{Se}, the latter systems being characterized by a smaller amount of clarity and an emphasis on lower sounds.

Thirdly, a separation is seen between the monophonic reproduction (2_M) and the two systems 4_{Me} and 6_{Se}, the former being negatively loaded in terms of the attributes *Broadness* and *Separability* and the latter being positively loaded in terms of the attribute *(no) Distortion*. The opposition between the attributes *Broadness* and *Distortion* illustrates the trade-off between spatial enhancement and timbral degradation for this type of mono-to-3D enhancement algorithms.

PCA of the music clip #2

The PCA model applied to the consensus data set of the music clip #2 explained 95.4% of the variation in the data with 3 components accounting respectively for 61.1%, 26.8% and 7.5% of the total inertia. Figures B.1 and B.2 in Appendix B show the score plots and loading plots of the PC 1–2 subspace and the PC 2–3 subspace.

The interpretation of these principal components appear to be globally similar to the data set of the music clip #1. The first component explains a large part of the variation in the data and correlates with attributes from the three perceptual groups defining the headphone vocabulary. The second component relates mainly to timbral aspects with the attribute *Tone Color*. The third component has a relatively small contribution but correlates clearly with the attributes *Broadness* and *Sense of Space*. From the two sets of PCA scores and their associated 95% confidence ellipses shown in Figure B.1, it can be observed that only the two systems 1_S and 7_{Se} were not discriminated (in a multivariate sense).

Combining the information from the score plots and loading plots, the following interpretation of this three-dimensional sensory map is proposed. Firstly, the monophonic reproduction (2_M) and the stereo enhancement systems 3_{Se}, 6_{Se} and 8_{Se} are opposed to the unprocessed stereo (1_S) and the two systems 7_{Se} and 4_{Me} in a global sense, that is, with a large number of attributes.

Secondly, the system 8_{Se} and the mono-enhancement system algorithm (4_{Me}) are separated from the other systems in terms of the timbral attribute *Tone Color* and, to some extent, the attribute *Clarity*.

Lastly, a separation similar to the music clip #1 result can be seen between the monophonic reproduction (2_{M}) and the systems 3_{Se} and 4_{Me} in terms of the attributes *Broadness* and *Sense of space* in one direction and the attribute *(No) Distortion* in the other direction.

PCA of the music clip #3

The PCA model of the third data set explained 92.4% of the variation in the data with 3 components while the other components accounted for a very small part of the total inertia. The score loading plots of the three first dimensions are shown in Figures B.3 and B.4 in Appendix B.

The interpretation of the principal components of this PCA model differs from the previous music clips in nature and order of importance. The first component explains 58.2% of the variation in the data and relates mainly to timbral attributes, i.e. *Tone color*, *Clarity*, and *Richness* although most of the attributes relate to this dimension. PC #2 recovers 25.3% of explained variance and relates to the attribute *Amount of echo* and to some extent *Broadness*, *(No) Distortion* and *Quality of echo*, which shares some similarities with the third dimension of the two previous PCA analyses. The third component explains 8.9% of the variation in the data and relates mainly to the attributes *Balance of Space* and *Balance of Sound*.

The position of the systems and their associated 95% confidence ellipses in the two score plots of Figure B.3 highlight a lower discrimination for this music clip as can be seen from the only three groups of systems clearly separated in the PC 1–2 subspace and the large overlap between ellipses in the PC 2–3 subspace.

A combined interpretation of this set of score plots and loading plots provides the following information. Firstly, the monophonic reproduction (2_{M}) and the stereo enhancement systems 3_{Se}, and 6_{Se} are opposed to the other systems in terms of attributes from the timbral and spatial groups. Secondly, the system 3_{Se} is characterized by a larger *Amount of echo*, more *Distortion* and less *Quality of echo* than other systems. Thirdly, the systems 6_{Se} and the monophonic reproduction (2_{M}) are characterized by less *Broadness* than other systems. Lastly, some level of separation between the monophonic reproduction (2_{M}) and systems 3_{Se} on one side and the systems (1_{S}) and 7_{Se} on the other side with respect to the attributes *Balance of Space* and *Balance of Sound*.

Similarities between the result of the three principal component analysis

To get a more global view of the spatial enhancement systems, a qualitative comparison of the three PCA results presented above was considered. The characteristics common to the three music clips can only be identified at a qualitative level in this type of comparison because the PCA space of the three music clips are not commensurable but in the present case however, this comparison turned out to be challenging. The only element clearly common to the three music clip was the grouping of three systems 1_{S}, 5_{Se} and 7_{Se}. An aspect common to the music clips #2 and #3 was also identified

relating to the opposition between 2_{M} and 3_{Se} in terms of the attributes *Broadness* and *Amount of echo*. However, it was found difficult to find reliable relationships between the other systems in these three sensory maps.

6.3.3 Global analysis by multiple factor analysis

In the preceding set of analyses, the sensory data under study was studied for each music clip separately but the qualitative assessment of the system differences across the three music clips from the PCA results was found to be difficult. To address this question of commonality and differences in results between the music clips, an analysis of the combined sets of data is considered below.

Two techniques at least are available for this type of analysis in the multivariate domain, that is, multivariate analysis of variance (MANOVA) and multiple factor analysis (MFA). MANOVA (Anderson, 2003) can be applied to the data set comprising the three arrays shown on the middle row of Figure 6.3 with the aim of assessing the effect of the factors `System`, `Assessor` and `MusicClip` and the different interactions while treating all the attributes as dependent variables simultaneously. MFA (Escofier and Pagès, 1988) can be considered as a direct extension of the PCA method to several groups of variables describing the same objects. In the present case, the set of variables are the three pre-processed two-way matrices from the PCA studies and the objects are the eight spatial enhancement systems under study. After an informal exploration of these two approaches, the author selected MFA for the present analysis because of its direct link to the PCA method already applied in this work and also because it offers useful tools to visualize the relationships between the objects, variables and groups of variables.

Presentation of the MFA technique

The principle of MFA is to perform a PCA on the merged pre-scaled group of variables. This pre-scaling is applied to each of the groups separately to ensure that they contribute equally to the PCA model. MFA can be described mathematically by considering each group of variables as a separate matrix X_i $(i = 1, ..., m)$ with n rows of objects and p_i columns. First, a weight λ_1^i is computed for the set X_i as the inverse of the first eigenvalue of the variance-covariance matrix $X_i' X_i$. Then, a PCA is applied to the merged matrix $X = (\lambda_1^1 X_1 | \lambda_1^2 X_2 | \ldots | \lambda_1^m X_m)$. Finally, the output of this PCA model can be represented visually by a set of graphical tools combining relevant information about X and X_i in a common PCA space. On the one side, the objects can be presented at two different levels on the same visual display, that is, the global scores of the matrix X and the superimposed 'partial' scores of the matrices X_i. On the other side, both the variables and the principal components of each separate PCA applied to X_i can be represented on a common correlation loading plot. Additionally, the L_g plot can be employed to visualize the link between the group of variables and each principal component of the matrix X.

The present analysis was performed with the MFA algorithm of the FactoMineR package (Lê *et al.*, 2008) running in the R software environment. The option 'centering' was selected, which put to zero the mean of each variable separately.

MFA results

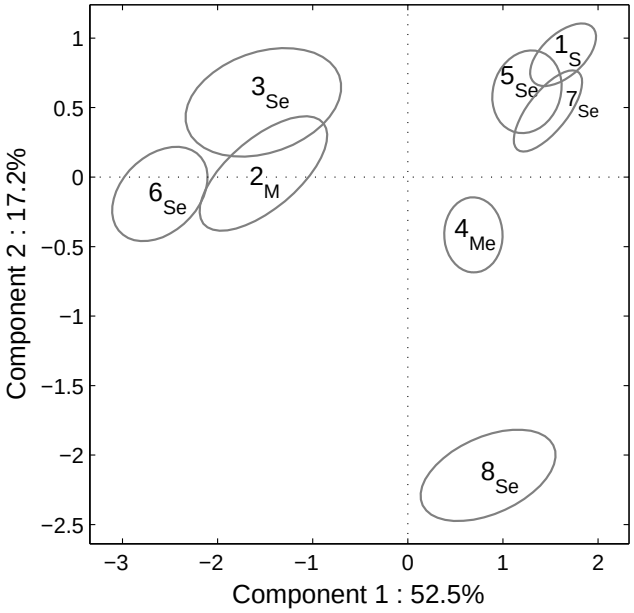
The MFA model applied to the three set of variables employed for the PCA study reported above explained 94.0% of the variation in the data with 4 components. The contribution of each component was 52.5%, 17.2%, 13.9% and 10.4% respectively and the other dimensions were found to have a smaller contribution (less than 4% explained variance) and were therefore not included in this analysis.

Starting the model interpretation with the objects, that is, the eight spatial enhancement systems, a visualization of the global score plots is illustrated Figure 6.8 for the two first dimensions (upper pane) and for the third and forth dimensions (lower pane). Ellipses around each system in these two plots represent the panel variability and have been derived with the approach presented by Pagès and Husson (2005), which follows the same principle as the PCA ellipse construction described in Section 6.3.2. A good level of separation between the systems can be seen overall, despite the presence of three groups of systems with overlapping ellipses. It can be seen from these two graphs that the unprocessed stereo (1_{S}) and the stereo enhancement algorithm 7_{Se} are not discriminated by the panel across the three experiments. However, all the other systems appear to be clearly separated in this 4-dimensional map, especially the systems 2_{M}, 3_{Se} and 4_{Se} on the third and fourth dimension. The partial score plots not presented here were also inspected and showed a relatively large variability between music clips for some systems, which illustrates the fact that the global scores conceal important differences between the music clips.

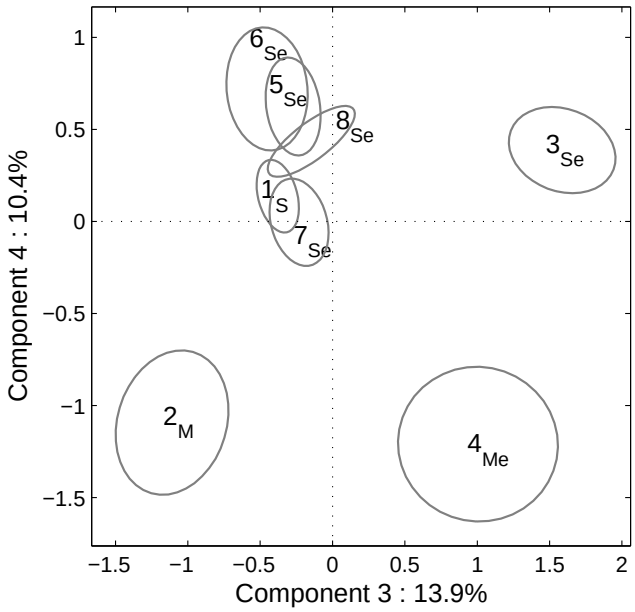
The correlation loading plot approach was adopted to visualize the relationship between the manifest variables (attributes) and the latent variables of the MFA model and to identify possible clustering of attributes within or across music clips. A separate correlation circle for each music clip was created to facilitate the visual interpretation and relevant perceptual directions were identified from this set of plots not presented here. A brief summary of the findings from these plots is provided next.

Most of the attributes were found to correlate positively with the first two dimensions but a perceptual direction was identified in the lower-right quadrant of the plot with similar timbral attributes from the three music clips, that is, the attributes *Tone color* and *Clarity*. Another group was found on the third dimension comprising the attributes *Broadness* and *Amount of echo* of the three music clips. Additionally, the fourth dimension was found to be characterized by a clear separation between attributes of the music clip #1 and the music clip #2.

The L_g plot shown in Figure 6.9 illustrates the link between the group of variables and the latent dimensions of the MFA model. It appears globally that the L_g index of all the groups of variables is high for the first dimension (left plot) but is relatively low for the other dimensions. In addition, important differences between music clips can be observed on the second and third dimensions indicating that the music clips #2 and #3 are better represented on the second and third dimensions respectively.



(a) MFA scores - PCs 1–2



(b) MFA scores - PCs 3–4

Figure 6.8: Score plot of the MFA dimensions 1-2 (upper pane) and 3-4 (lower pane). Ellipses around the system scores represent 95% confidence regions based on assessor variability.

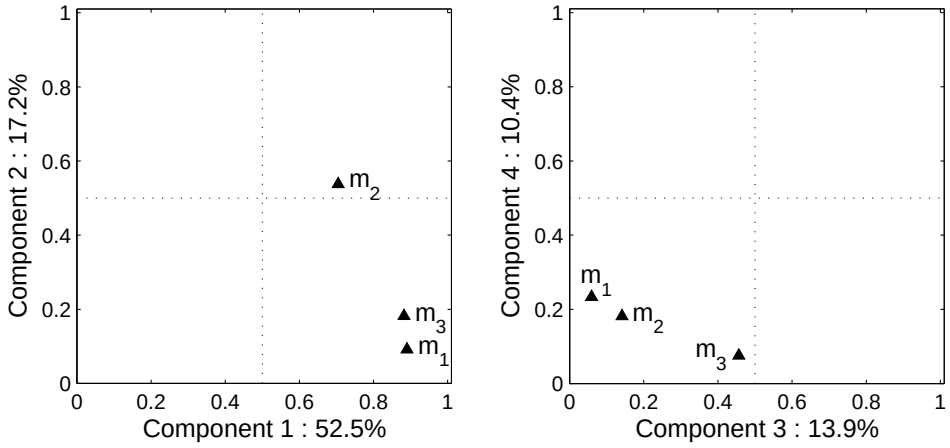


Figure 6.9: L_g plots illustrating the contribution to the four MFA dimensions of the three music clips.

Summary of the global analysis

The combined analysis of the three music clips applied with the MFA technique allowed to derive a sensory map of the eight spatial enhancement systems that is common to the three music clips and to assess the differences between music clips. A good discrimination between the systems was highlighted at a global level and the interpretation of the sensory map can be summarized as follows.

- the systems 2_{M}, 3_{Se} and 6_{Se} are clearly separated from the other systems in global terms, that is, with a large number of attributes,
- the system 8_{Se} is characterized by more *Tone color* and *Clarity* than the other systems for the three music clips,
- the systems 3_{Se} and 4_{Me} are characterized by more *Broadness* and *Amount of echo* than the other systems.

6.4 Discussion

The attribute rating experiment presented in this chapter unraveled the important perceptual characteristics of spatial enhancement systems for headphone reproduction. The spatial and localization attributes of the consensus vocabulary employed in this test were found to discriminate clearly the mono reproduction from the unprocessed stereo reproduction. However, it was also found that the stereo enhancement systems 3_{Se} and 6_{Se} are perceptually closer to the monophonic down-mixed sound reproduction than the unprocessed stereo. The mono-to-3D algorithm (4_{Me}) included in this experiment showed some enhancement with respect to the monophonic material, but only in terms of spatial attributes and not for the localization attributes, as expected. The timbral attributes *Tone color*, *Clarity* and *Richness* were also found to discriminate the original stereo samples from the stereo enhancement systems and the mono-to-3D algorithm. It should be noted that these timbral aspects were associated to a quality degradation in the set of preference experiments reported in Chapter 3.

The result of this experiment also shows some limits of the headphone vocabulary developed in this work. Overall the level of discrimination achieved by the consensus panel was lower than expected and a very large correlation between the attributes was found. It is expected that further training of the assessors would have increased the discrimination ability of the assessors and the agreement of the panel. Regarding the second issue of large attribute correlation, the reason might be that the set of attributes employed for this experiment was not created directly for the stimulus set under study but was the result of a descriptive analysis work aiming at a generic description of headphone sound. A refinement of the attributes to fit better the scope of the stimuli evaluated in this study, for example through additional panel discussions supported by a training in attribute scale usage, might have been beneficial for this type of consensus vocabulary application.

Chapter 7

Overview of individual vocabulary methods

7.1 Introduction

The group of individual vocabulary (IV) methods introduced in Chapter 2 belong to the class of descriptive sensory analysis techniques based on verbal elicitation and letting each assessor develop his or her own set of sensory descriptors. The IV approach represents an alternative to the consensus vocabulary (CV) methods found in the same class of verbal elicitation methods. While the CV approach was covered in depth in Chapters 4 to 6 of this thesis, Chapters 7 to 9 will focus on the IV approach with a similar structure including a review of this group of test methods in the present chapter, the presentation of a novel procedure for running IV experiments with inexperienced sensory assessors in Chapter 8, and the application of this type of IV technique to the set of headphone sound stimuli already encountered in Chapter 6.

This chapter is organized as follows. A broad overview of the IV approach is provided in Section 2 including a literature review on existing techniques and a discussion on the commonalities between these procedures. Aspects relating to the analysis of data resulting from IV experiments, later referred to as 'individual vocabulary data', are developed in Section 3 and the issue of IV data validity is addressed in Section 4.

7.2 Historical background of individual vocabulary methods

CV methods and IV methods belong to the class of descriptive analysis techniques and therefore they both assume that (trained) subjects can break down their perception into its constituting elements and can give a meaningful verbal description of these perceptual components. In Chapter 4, a description of the principles for developing a CV was provided and several CV techniques were reviewed. These advanced experimental procedures are designed to ensure that assessors agree on the selection of a set of sensory descriptors and use the attributes in a consensual way. One major

characteristic of IV methods is that they circumvent totally this matter of consensus. More specifically, the term 'individual' highlights the idea that assessors build their own set of descriptors without interacting with the rest of the panel. This idea is illustrated in the review of IV techniques provided next.

7.2.1 The three major individual vocabulary methods

The group of IV methods presented in Section 2.3.2 includes three important techniques thoroughly documented in the literature which are Free-Choice Profiling, the Repertory Grid Method and Flash Profile. These three techniques are presented below to illustrate the different approaches encountered in practice and to give insight to the chronological evolution of the reasoning behind IV after an application period of about 30 years. It should be noted that other techniques, although less documented, do exist. For example alternative attribute elicitation approaches have been exploited like the 'natural grouping' technique in which assessors first create groups of stimuli based on similarity and then verbalize the perceived difference(s). The use of hybrid versions of these three major techniques can also be conceived as will be shown in the next chapter. Using an analogy to the generic (consensus) descriptive analysis discussed in the overview of CV methods presented in Chapter 4, the similar term of 'generic (individual) descriptive analysis' might be applied to account for the group of techniques combining elements of the three methodologies to be described next. It should also be mentioned that the review considered in this section concerns only descriptive sensory analysis. Free-Choice Profiling and the Repertory Grid Method have also been exploited in consumer studies but a clear distinction is made in this thesis between the evaluation of the sensory characteristics of a stimulus set by either expert sensory assessors or naive subjects on the one side and the evaluation of product characteristics relating to attitudinal aspects on the other side. The latter measurement belongs to the 'affective domain' as described in Chapter 2 and is therefore not covered in this chapter.

Free-Choice Profiling (FCP) is the method that received the most attention in the field of sensory science. Introduced by Williams and Langron (1984), this technique has been tested on various food categories by British sensory scientists in the 1980's, e.g. Williams and Langron (1984) on wines, Williams and Arnold (1985) on coffee, Marchall and Kirby (1988) on cheese, Guy *et al.* (1989) on whiskey and McEwan *et al.* (1989) on chocolate. Early publications presented FCP as an alternative approach for sensory profiling that could overcome some shortcomings of consensus techniques such as QDA. Assuming that subjects differ mainly in the way they describe sensory characteristics and not so much in the way they perceive them, FCP allows assessors to elicit and quantify these characteristics with their own vocabulary. Consequently, the effort used for panel training is considerably reduced because the difficult and time-consuming step of concept alignment, that is, the process in which assessors adjust to an agreed common set of descriptors, is not required in this case. The application of FCP has been initially illustrated with assessors experienced in sensory profiling but the method has also been applied with inexperienced assessors in many consumer studies (see Jack and Piggott (1991) for a comprehensive review). The at-

tribute rating of the samples is generally made using a monadic sequential paradigm similar to the method employed in QDA, i.e., one stimulus is presented at the time and is evaluated for all attributes. An additional specificity of FCP relates to the use of a data analysis procedure known as Generalized Procrustes Analysis (GPA; see Gower, 1975), which is specifically designed to handle the individual profiles resulting from this type of sensory evaluation with the aim of producing a consensus profile. Note that the meaning of the term 'consensus' used in this context differs from the semantic agreement obtained with the CV approach. Dijksterhuis and Gower (1991) proposed to use the term 'GPA group average' to acknowledge the fact that such an average might conceal a wide range of difference of viewpoint. Although alternative analysis techniques exist, GPA has been closely linked to the development of FCP and remains today the favored method for the treatment of this type of sensory profiling data. Hence, the GPA technique will be described more thoroughly in the latter part of the present chapter.

The Repertory Grid Technique (RGT) did not originate from the food industry but was developed by Kelly (1955) in the field of psychology during the 1950's. The main idea of this technique already mentioned in Chapter 2 is to get subjects to define their own constructs by describing the ways in which elements and their associated meanings vary. The Repertory Grid has also been applied to the field of market research (Frost and Braine, 1967) and the first applications of this technique in sensory science occurred in the early 1980's. Olson (1981) introduced the idea of using this technique to elicit the characteristics of food products. In this context, a structured approach of triadic presentation was applied, which allowed each assessor to state the characteristic for which two of the samples were similar to each other and different from a third one. After a number of triads, an individual set of constructs defining the product space was obtained and could then be applied for the evaluation of all the products. Different types of analysis have been exploited to study individual and multiple grids resulting from RGT experiments, including non-parametric factor analysis (Kelly, 1963), principal component analysis (Slater, 1977) and cluster analysis (Shaw, 1980). In the field of food science, Thomson and McEwan (1988) recognized that the constructs elicited by the RGT are in fact analogous to the attributes obtained in FCP and they proposed to apply GPA to the resulting set of individual sensory profiles. Studies comparing RGT and FCP have been reported in the literature and some authors, e.g. McEwan *et al.* (1989) or Piggott and Watson (1992), favored RGT claiming that this more structured approach is beneficial to inexperienced assessors during the IV development process. More recently, RGT has also found applications in the field of audio with some adaptations. For example, Berg (2005) developed a software tool to facilitate the elicitation, rating and analysis steps of the RGT procedure, which employed a comparative grading of all stimuli for a given construct. Choisel and Wickelmaier (2006) selected a similar approach and they compared the use of either diads or triads in the elicitation phase.

Flash Profile (FP) is a more recent IV technique introduced in the field of sensory science by Sieffermann (2000). This methodology combines the individual elicitation approach of FCP with a comparative evaluation technique. The comparison of all

products during the descriptive analysis process is claimed to remove the need for a phase of product familiarization and a phase of individual training with the attributes. This reduces therefore the whole process to the sensory elicitation phase and the attribute rating test. In addition, FP employs assessors familiar with descriptive analysis to ensure that discriminant and non-hedonic attributes can be generated in a limited time. As a result, a relative sensory positioning of the products can be obtained in just one to three sessions with this technique. Several studies comparing FP to a generic consensus profiling method have been reported. Dairou and Sieffermann (2002) tested a set of products with large sensory differences and obtained a comparable sensory map with the two approaches while Delarue and Sieffermann (2004) selected two more similar sets of products and found the same results in only one of the two experiments they performed. These authors noted however that the semantic interpretation of the product sensory space was more challenging with FP than with the CV approach. An additional limitation of Flash Profile relates to the comparative evaluation procedure adopted. This approach requires all the products to be available simultaneously and restricts the number of products that can be compared due to issues of assessor fatigue. A recent publication by Tarea *et al.* (2007) illustrated however that FP is still applicable with a relatively large set of samples, that is, 49 food products in this case. Nevertheless, the descriptive analysis task was found to be difficult and time-consuming for assessors.

7.2.2 Chronological view of individual vocabulary methods

It appears from this short review that the IV approach is a relatively new tool in sensory science in comparison to the CV approach. Early applications of FCP and RGT in the food industry only occurred in the 1980's while consensus methods (often referred to as 'conventional' profiling methods) like the Flavor Profile method and QDA were developed respectively in the 50's and 70's. It should be emphasized that IV techniques have been developed in the continuity of existing sensory methods. They follow the principles of descriptive sensory analysis described in Chapter 5, which are 1) employing a panel of (trained) assessors rather than a single expert, 2) developing an objective description of products based exclusively on sensory properties, and 3) performing quantitative evaluations with (reliable) scales. However, these techniques introduced a clear rupture from the CV methods in the way the descriptive attributes are generated by the sensory panel and handled by the experimenter.

Considering retrospectively the introduction of FCP in the field of food science, it is worth mentioning that the group of early publications listed above presented this method as an alternative to consensus methods with rather strong claims. For example, the comparison of conventional (i.e., consensus) profiling, FCP and similarity scaling methods for coffee aroma evaluation reported by Williams and Arnold (1985) highlighted the advantages of FCP over consensus techniques in terms of precision, sensitivity and ease of application and they stated in conclusion that "*there are many instances, therefore, where it* (FCP) *could provide a more appropriate, faster and cheaper alternative to conventional profile assessments*". Strong criticism towards FCP followed, as illustrated for example in the review of this technique by Stone

and Sidel (1993)[1]. The reasons usually invoked against this new approach concerned the issue of assessor reliability when no training is provided, the heavy manipulation applied to the data through GPA to arrive to a consensus (Huitson, 1989) and the risk of interpretation bias due to the idiosyncratic nature of individual attributes. It should be mentioned however that some of these issues have been addressed in more recent literature. For example, statistical procedures to validate the consensus configuration obtained from GPA have been proposed based on monte-carlo simulations using random data (King and Arents, 1991) or permutation techniques (Wakeling *et al.*, 1992).

The comparative studies of RGM and FCP mentioned above highlighted the fact that despite the differences between the elicitation process employed in these two methods, the resulting individual vocabulary is essentially of the same nature, the attribute rating phase is more or less identical and the statistical processing of the resulting data can be applied with the same tools, e.g. GPA. It can be noted that a wider range of elicitation methods exist as presented by Steenkamp and Trijp (1997) or Bech-Larsen and Nielsen (1999) in the field of consumer market. It would appear that in practice the selection of an elicitation method for an IV experiment is relatively open and can be adapted to the level of experience of the sensory assessors employed for the task.

Considering at last the recent introduction of Flash Profile, which happened about 20 years after the development of FCP, a shift in reasoning can be observed. Indeed, the emphasis in FP is to produce a perceptual mapping of a product space using a rapid sensory procedure and the method has been presented more as a sensory tool complementing conventional techniques rather than replacing them. For example, FP has been applied as a preliminary phase before a more thorough sensory study by Dairou *et al.* (2003) and Tarea *et al.* (2003).

7.2.3 Summary on individual vocabulary methods

To conclude this introductory section, a generic description of the IV approach is given in light of current application trends. This sensory evaluation methodology can be seen as a less involved technique than the CV approach and is organized in the three following main steps: 1) an IV development using a more or less structured elicitation method depending on the sensory expertise of the assessors, 2) an individual attribute rating phase and 3) an experimental interpretation of the resulting sensory profiles at an individual level and (more importantly) at a panel level with the dual goals of deriving a global sensory map of the set of stimuli under study and interpreting its meaning from a semantic point of view. Beyond the practical benefit that the IV approach brings by avoiding the process of consensus vocabulary development, the use of individual vocabularies also offers some advantages in terms of methodology. Indeed, this approach introduces minimal bias in the sensory evaluation process because subjects are given the freedom to choose a vocabulary that fits best their perception and sensitivity instead of being forced to use a common vocabulary. Additionally, the combination of all the individual vocabularies enriches the global sensory description

[1]A reply to the criticism stated in this textbook can be found in Dijksterhuis and Heiser (1995).

and offers a means to test the importance of specific sensory aspects at a panel level. Nevertheless, these advantages are counter-balanced by the added difficulty arising when a global picture has to be drawn from these multiple individual descriptions. This ultimate stage represents the challenging side of the IV approach and is developed further in the next section.

7.3 Analysis of individual vocabulary data

The overview presented above highlighted the idea that IV methods circumvent the issue of panel consensus by allowing each assessor to profile the stimuli under study with his or her own set of sensory descriptors. This approach has some methodological benefits as discussed earlier but it also introduces issues not encountered in CV experiments because the attributes selected by different assessors are not directly commensurable. Firstly, testing formally the semantic validity of the individual attributes is a difficult matter. Secondly, the task of finding relationships between the attributes of different assessors either at a qualitative or a quantitative level requires additional efforts from the experimenter and is prone to interpretation bias. Thirdly, the derivation, validation and interpretation of a global sensory map, i.e., a description representing the sensory panel view, can be a complex matter especially for the semantic characterization of stimulus differences as discussed by Dairou and Sieffermann (2002) or Delarue and Sieffermann (2004). The first part of this section will address the two first aspects while the second part will focus on the third aspect which relates to a specific research topic of sensometrics looking at methods to derive a 'common' representation[2] from a set of individual sensory profiles and to study individual variations around this global description.

7.3.1 Specifics of individual vocabulary data

The review of the CV development process presented in Chapter 4 underlined the importance of panel agreement regarding the meaning and rating of the sensory descriptors. This effort effectively ensures that the resulting sensory description is consensual, reliable and easy to interpret by the experimenter owing to the definitions associated with the attributes. On the contrary, the validity of the sensory descriptions produced by an IV panel has to be demonstrated a posteriori by the experimenter or data analyst due to the heterogeneity of the individual vocabularies.

Gaines and Shaw (1993) proposed a methodology for eliciting knowledge from multiple experts in which they address the issue of comparing individual vocabularies. Although the domain of application of the system they developed is more general than our specific descriptive sensory analysis case, the elicitation process is comparable. These authors use RGT to elicit from subjects two types of information about entities (the stimuli under study): the 'distinctions' made between entities and the

[2]Such a representation is also referred to as 'consensus', '(group-) average' or 'compromise' in the literature.

'terms' used for these distinctions[3]. When comparing the conceptual systems elicited by two experts, the couple term/distinction can give rise to four scenarios, namely consensus, conflict, correspondence and contrast. *Consensus* arises if two different conceptual systems assign the same term to the same distinction, *conflict* arises if the same term is assigned to different distinctions, *correspondence* arises if different terms are assigned to the same distinction and *contrast* arises if the conceptual systems assign different terms to different distinctions. This interaction between terminology and distinctions is illustrated in Figure 7.1. Gaines and Shaw (1993) highlighted the importance of recognizing each of these situations to better measure the level of commonality and idiosyncrasy between individual conceptual systems and they implemented an iterative elicitation process to identify these four scenarios.

The IV data structure applies well to the terminology/distinction relationship presented above because it comprises individual vocabularies (qualitative data) and associated ratings (quantitative data). In this respect, it is useful to keep the scenarios illustrated in Figure 7.1 in mind when analyzing and interpreting this type of data. Two complementary routes exist to compare the sensory descriptions produced by the assessors of an IV panel. A bottom-up interpretation strategy can be considered starting from the individual data and seeking a global view or a top-down approach can be considered in which a backward semantic interpretation of the individual features is performed directly from a global sensory map representing an average view of the panel.

[3]In Kelly's Personal Construct Psychology (1955), a construct is a basis for making a distinction, that is, a dichotomous reference axis. So, the term 'distinction' can be compared to the term 'attribute' defined in Chapter 2.

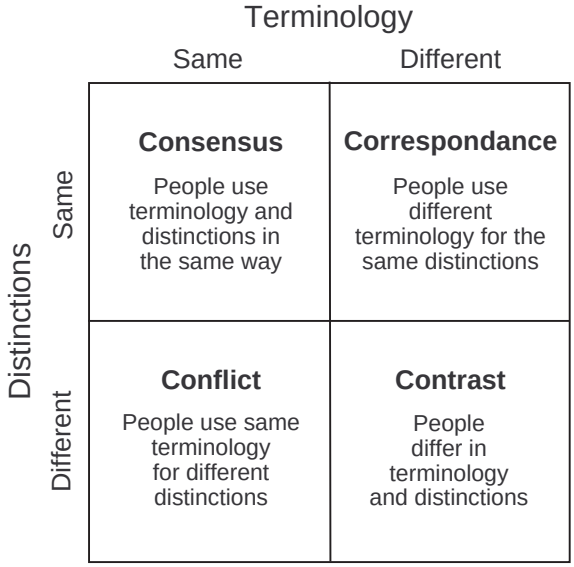


Figure 7.1: Representation of consensus, correspondence, conflict and contrast between distinctions (after Gaines and Shaw, 1993).

In the former approach, the focus is usually placed on the terminology, for example by looking for similar terms across subjects or by creating groups of semantically associated descriptors. Verbal protocol analysis, a method developed by Samoylenko *et al.* (1996) for the qualitative analysis of free verbalization data, has been applied for this purpose in the IV context by, e.g. Berg and Rumsey (2000) and Neher *et al.* (2006). Irrespective of the technique adopted, it should be noted that quantitative data can be exploited to identify the scenarios of *consensus* and *conflict* described above. Additionally, the strategy of finding relationships between *distinctions* based on hierarchical cluster analysis or PCA of one or several sensory profiles can also be applied to identify *correspondence* scenarios and measure the level of *contrast* between individual vocabularies.

In the latter approach the complete set of individual sensory profiles is considered at once and with the help of a specific type of statistical analysis (to be discussed later in this section), a global sensory map is derived from which it is possible to compare individual descriptions. The results can then be interpreted in terms of terminology/distinction relationships based on the co-variation of individual attributes in the derived common sensory space. This allows to identify to some extent the four scenarios presented in Figure 7.1 but it should be noted that there is formally no way to disentangle the *correspondence* scenario from the case where two different distinctions co-vary in an IV data set.

7.3.2 'Group' analysis of individual vocabulary profiles

Although the three IV techniques described in the previous section vary somewhat in terms of vocabulary development process, it was highlighted earlier that the quantitative data generated in this type of experiments fit usually the same format. As each individual evaluates the same products with his own set of attributes, the resulting group of individual sensory profiles can be described as a multivariate data set comprising K matrices (one for each assessor) with N rows (the objects representing the samples under study) and P_i columns (the variables representing the attributes employed by the assessor i). Note that a single matrix is often referred to as a 'configuration' of objects in this context. The fact that the number P_i depends on the size of the IV and that no relationship can be assumed between the attributes of different assessors implies that a direct data averaging over the assessors as applied to CV data is not feasible in this context[4]. However, as the assumption in the IV approach is that assessors perceive (to some extent) the same sensory characteristics, the individual configurations are expected to share some similarities and the aim therefore is to derive an interpretable description of the differences between objects at a panel level. This can be achieved with multivariate analysis techniques capable of handling the individual nature of these sensory profiles. Generalized Procrustes Analysis is the example the most commonly encountered in sensory science for this type of analysis and it will therefore be thoroughly reviewed below to illustrate the underlying principles of IV data analysis. However, other techniques can also be applied to this category of

[4]The type of data set encountered in the IV context is often referred to as a K-set (e.g. Dijksterhuis, 1996) by opposition to the 3-way data set usually encountered in the CV context where variables in different sets are assumed to be commensurate.

sensory profiling data and several of them will be introduced in the last part of this section.

Generalized Procrustes analysis

The term 'Procrustes' was first employed by Hurley and Cattell (1962) to describe the process of matching two matrices. The Procrustes problem can be formulated by considering a matrix of N objects by M variables as a geometrical configuration. These N points lie in a M-dimensional space (a set of orthogonal Cartesian axes representing the measurement variables) and the distances between the points in this space reflect relations between the objects. The idea of (orthogonal) Procrustes analysis is to apply a set of geometrical transformations to one of the configurations to obtain an optimal match to the other one under the constraint that the relative distance between the samples of the transformed configuration is preserved. The geometrical transformations allowed in this procedure include the translation, the rotation/reflection[5] and possibly the isotropic scaling of the configuration. Through this process, the same objects of the two different configurations are brought as close as possible to each others in the multidimensional space under the given constraint. The resulting fit can be measured with the Procrustes statistic, which is defined mathematically as the sum of squares of the Euclidean distances between the two configurations.

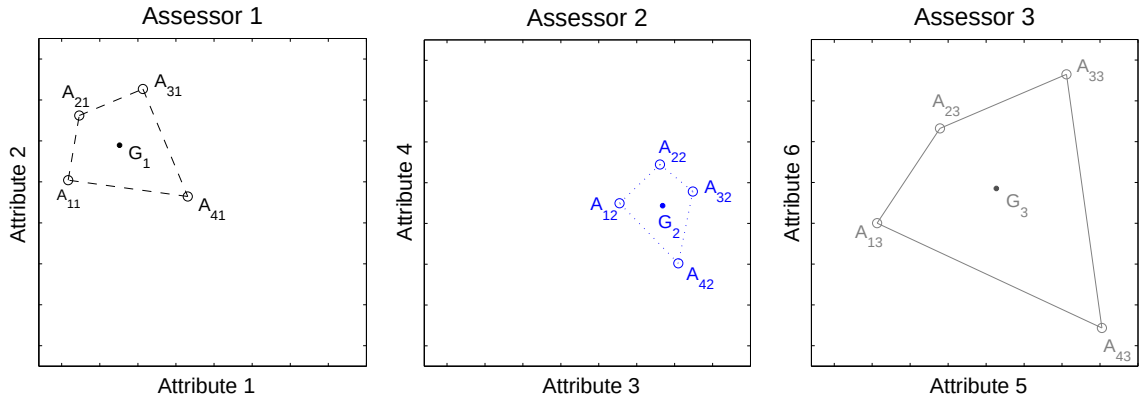
A generalization of the Procrustes idea to a set of configurations with possibly different numbers of columns was developed in the 1970s. Kristof and Wingersky (1971) first introduced the idea of a common configuration and later Gower (1975) gave a complete formulation of the GPA technique. In this procedure, individual configurations are iteratively transformed by translation, rotation/reflection and isotropic scaling with the aim of minimizing the Procrustes distance between each individual configuration and an average configuration corresponding to the mean of the individual transformed configurations. More details about the mathematics associated with Procrustes analysis and GPA[6] can be found in Gower and Dijksterhuis (2004).

GPA can be run with several software packages including the Senstools for Windows v3.2 (2004) or the XLSTAT add-in for Microsoft Excel (2006). An implementation of GPA handling missing values can also be found in the FactoMiner package of the R software (R Development Core Team, 2003). The GPA routine employed in this thesis was implemented in Matlab by the author based on the algorithm presented in Chapter 9 of Gower and Dijksterhuis (2004). This Matlab routine can be found in Appendix C.

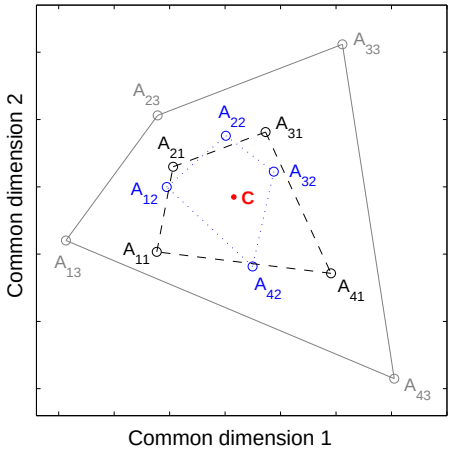
The application of GPA to IV data has been reviewed in several publications including Arnold and Williams (1986) and Dijksterhuis (1996). When applying this type of analysis technique to sensory profiling data, the existence of a 'true' configuration

[5]The geometrical transformation represented by the orthogonal matrix considered in this case depends on the sign of its determinant. The absolute value of this determinant is always 1 and the cases +1 and -1 corresponds respectively to a rotation and a reflection (Gower and Dijksterhuis, 2004).

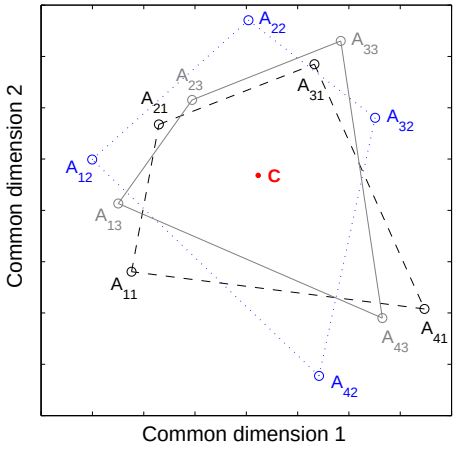
[6]It should be noted that GPA is referred to as 'generalized orthogonal Procrustes analysis' (GOPA) in recent publications (see e.g. Gower and Dijksterhuis, 2004; Arnold *et al.*, 2007) to clarify the nature of the transformation matrix used in this procedure. Indeed, orthogonal Procrustes is only one case of the more general Procrustes problem which also includes projection and oblique Procrustes.



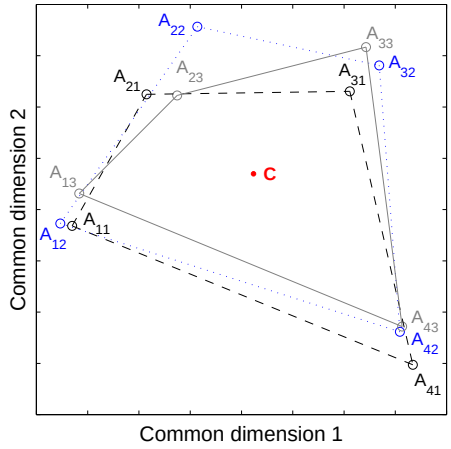
(a) Original configuration of three assessors. The point A_{ik} represents the object i of the assessor k and G_k is the centroid of the individual configuration of the assessor k. Note that the individual attributes are assumed to differ across assessors but should be perceptually related in this example.



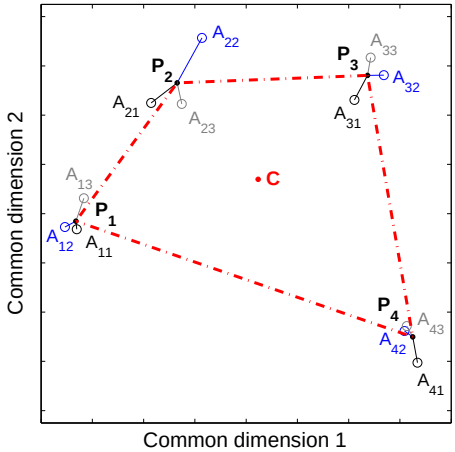
(b) Centering of configurations to the centroid C set as the origin of the common space.



(c) Isotropic scaling of configurations.



(d) Rotation/reflection of configurations.



(e) Final average configuration and associated assessor variation around each object P_i.

Figure 7.2: Geometrical illustration of the GPA procedure (adapted from Dijksterhuis and Gower, 1991).

specifying the distances between the set of objects under study is assumed and each individual sensory profile is supposed to reflect this configuration to some extent. The transformations operated on individual configurations seek to compensate for characteristic differences in the way assessors use their attributes. Figure 7.2 adapted from Dijksterhuis and Gower (1991) illustrates the three steps of this procedure for a simulated sensory data set comprising four objects evaluated by three assessors using their own set of attributes. For the sake of visual clarity, it is assumed in this example that the assessors employed only two attributes but the procedure applies also to vocabularies with a larger size. The three original configurations are presented on a separate plot in Figure 7.2(a) to highlight the idea that individual attributes are not commensurable in this type of data set. Large differences in level and range of scoring can be observed between the three assessors. The translation step of GPA aims at removing differences in level of scoring as illustrated in Figure 7.2(b). Note that the semantic meaning of the scales is lost when the three configurations are superimposed on the same graph. The axes represent now a common set of dimensions with the size of the largest individual vocabulary of the sensory data set. The isotropic scaling step addresses the issue of differences in scale usage between assessors by allowing a stretching or shrinking of the entire individual configurations as demonstrated in Figure 7.2(c). The rotation/reflection step aims at aligning attribute descriptions to obtain as similar configurations as possible in the common multidimensional space (Figure 7.2(d)). The final individual transformed configurations are obtained after a number of iterations of scaling and rotation steps and an average configuration can be derived as illustrated in Figure 7.2(e). The remaining individual variation for each object is exhibited in this graph as the distance between the point A_{ik} of an assessor and the point P_i representing a 'panel average' in the sense of GPA.

It should be noted that a specific set of assumptions is made for each of these three transformations. The translation step is easy to justify as it simply puts the centroid of all individual configurations to zero. This leads to a relative product description as commonly found in multivariate analysis techniques including the pre-processing step of variable centering (i.e. setting the mean of each column to zero). The isotropic scaling step is more questionable because it assumes that a given assessor uses a similar range for all the attribute scales, which is not always verified as illustrated by Næs (1990). However, accounting for range differences on an individual attribute basis would complicate the matching procedure. Such an anisotropic scaling technique has been discussed e.g. by Gower and Dijksterhuis (2004) and its application to sensory data has been presented by Næs (1990) and Hanafi *et al.* (2004) but it is applied very seldom in practice. Finally, the rotation/reflection step is the most radical transformation of the GPA procedure in the sense that it does not make any assumption about the semantic meaning of individual attributes but merely adjusts the relative position of the objects under the given geometrical constraints. It is only a posteriori that derived correlations between the attributes of different assessors can be assessed from a semantic viewpoint.

One aspect of the isotropic scaling transformation discussed in the literature concerns the use or not of an individual configuration scaling prior to the GPA procedure. Arnold (1992) highlighted the fact that the original isotropic scaling in GPA includes

both an overall size adjustment and a differential weighting of each individual configuration. He proposed to apply a pre-scaling of each individual matrix to a unit trace to handle the configuration size adjustment prior to GPA and noticed that the isotropic scaling subsequently included within GPA produces the same solution while giving a measure of agreement with the rest of the panel for each assessor. Note also that other scaling approaches can be adopted, such as the P_k-scaling described by Dijksterhuis (1996) including an additional data scaling for each attribute separately.

The main output of the GPA procedure is a group average configuration and a set of transformed individual configurations. All these configurations have the same dimension and lie in a common multidimensional space but the meaning of these dimensions has to be interpreted indirectly from the original individual attributes. Usually, a PCA is applied to the average configuration to represent the object map in a space of lower dimensionality (i.e. the subspace defined by the two or three first dimensions) and the individual attributes are then examined in this space. The original variables can be represented directly in the form of loadings or through their correlation with the principal components as described by Dijksterhuis (1996). The biplot approach proposed by Gower and Hand (1996) can also be exploited to interpret visually the object-attribute relationships of the average configuration obtained from the GPA procedure as illustrated by Arnold *et al.* (2007). Additionally, the transformed individual configurations can be superimposed on the PCA score plot by a projection technique to illustrate individual variability around the group average scores on the two or three first principal components[7]. The residual variation per assessor in the remaining dimensions also gives an estimation of the global fit of each individual configuration to the group average in the low-dimensional representation of interest.

The analysis of variance framework associated to GPA referred to as 'Procrustes analysis of variance' (PANOVA Gower, 1975) is also a useful tool to assess the significance level of the different GPA steps. But more importantly, techniques to assess the validity of the group average configuration should be applied to ensure that the GPA procedure effectively models relevant information in the data and not just noise. Resampling methods can be employed to test the statistical significance of the GPA solution using either monte-carlo simulations with random data (King and Arents, 1991) or a random permutation technique (Wakeling *et al.*, 1992).

Alternative multivariate data analysis approaches

The combination of FCP and GPA has been widely used in the field of sensory science following the seminal work by Williams and Langron (1984) and GPA remains today the dedicated technique for IV data analysis. However, alternative approaches have also been presented in the literature for this type of analysis including e.g. PCA, MFA and STATIS. During the course of this thesis work, the author explored and compared these techniques. Through this exercise, it has been possible to put GPA into a wider perspective and to better understand the meaning and the limits of the transforma-

[7]Note that the score of an object in the PCA applied to the group average configuration represents the barycenter of the set of projected individual scores and the dispersion of the individual objects reflects therefore the level of agreement amongst assessors for this object.

tions associated with this classical approach. A brief presentation of four analysis techniques applicable to IV data is proposed below and the relationships between these different approaches in terms of individual difference handling are discussed.

Principal component analysis is the simplest form of multivariate analysis that can be applied to IV data. Considering each individual sensory profile as a separate matrix X_i with n rows of objects (the samples under study) and p_i columns of centered variables (the attributes of the assessor i), this technique gathers horizontally the full set of m matrices to form a matrix $X = (X_1|X_2|\ldots|X_m)$ of size $n \times \sum p_i$. Applying a PCA to such a matrix yields a set of principal components with scores common to the individual sensory profiles[8]. This approach has been described by Kunert and Qannari (1999) who proposed an isotropic pre-scaling of the individual sensory profiles to a unit sum-of-squares (i.e., $trace(X_i'X_i) = 1$) in order to impose an equal contribution to the model for each individual sensory profile irrespective of the number of attributes.

Multiple factor analysis was introduced by Escofier and Pagès (1988) for the joint analysis of several sets of variables relating to the same objects, which is precisely the format of an IV data set. This technique already exploited in Chapter 6 is similar to a PCA applied to the merged set of variables as described above but it applies a specific pre-scaling to each set of variables in order to balance the contribution of these sets in the analysis. The weight applied to the set X_i of centered variables is the inverse of the first eigenvalue of the variance-covariance matrix $X_i'X_i$. Graphical tools similar to PCA are available to visualize the common objects and individual attributes and to assess the relationships between each group of variables and the different latent components of the model.

STATIS[9] is a three-way multivariate analysis method introduced by L'Hermier Des Plantes (1976) (see also Lavit *et al.* (1994) for a general presentation and Schlich (1996) for an application to sensory data analysis). Unlike the previous methods, STATIS works from the association matrix which is defined as $W_i = X_iX_i'$ where X_i has centered variables. The dimension of the square matrix W_i depends on the number of objects but not on the size of the individual vocabulary and it contains all the information about the multidimensional differences between objects found by the assessor i. Based on the RV coefficient Escoufier (1973); Robert and Escoufier (1976) between these individual association matrices, an individual scaling factor proportional to the level of agreement between a given assessor and the rest of the panel can be derived. The weighted mean of the W_i defines the STATIS compromise and a map of the common objects and individual attributes can be obtained from the principal components of this common association matrix.

PARAFAC2 was introduced by Harshman (1972) and belongs to the class of multiway analysis techniques described e.g. by Smilde *et al.* (2004). The PARAFAC1[10] model can be considered as one generalization of PCA to higher order arrays. The idea is, in the case of a three-way array for example, to obtain a decomposition into a set of trilinear elements for each latent component. A consensus sensory profile

[8]Note that this unfolding technique when applied to CV data is referred to as the 'common scores' Tucker-1 model described e.g. by Brockhoff *et al.* (1996).

[9]This French acronym can be translated to 'Structure of three-way data set in statistics'

[10]PARAFAC stands for 'parallel factor analysis' and is named PARAFAC1 here to distinguish it from the PARAFAC2 method.

comprises the three modes `Assessor`, `Sample` and `Attribute` and can therefore be analyzed with PARAFAC1. The outcome of this model includes a set of sample scores and attribute loadings as found in a usual PCA but also a set of assessor loadings for each latent component. The PARAFAC2 method works on a similar principle but it allows variables in one mode to vary across another mode and similarly to the STATIS method it uses the association matrix $W_i = X_i X_i'$ for the changing slab[11], which corresponds to the matrix `Sample` by `Attribute` in an IV data set. In the direct fitting model proposed by Kiers *et al.* (1999) (see Appendix E), the output of PARAFAC2 includes for each latent component a vector for the mode `Assessor`, a vector for the mode `System` and one vector specific to each assessor for the mode `Attribute`. Note that the three-way analysis method proposed by Qannari *et al.* (2000) under the name 'Common Components and Specific Weights Analysis' shares a similar principle and has been applied to sensory data. In Chapter 9, a more detailed presentation of PARAFAC2 will be provided in the context of a practical application to a four-way IV data set.

The short review presented above gives an illustration of the alternative approaches to handle variations between individual sensory profiles. It should be mentioned first that all these multivariate data analysis techniques work with centered variables. A less intuitive idea to highlight from this review concerns the rotational freedom given to the individual configurations in all the analysis methods presented above although this is only directly exhibited in GPA and PARAFAC2. In the case of PCA and MFA, the horizontal merging of the individual matrices allows the dimensions common to all data sets to emerge irrespective of the original orientation of each individual configuration. In fact, the rotation of an individual matrix has no effect on the outcome of this type of PCA as demonstrated by Kunert and Qannari (1999). Similarly, STATIS handles indirectly the aspect of individual configuration rotation through the use of the association matrix. In the PARAFAC2 case, the iterative algorithmic procedure proposed by Kiers *et al.* (1999) includes an explicit rotation step and the final rotation matrix is exhibited from the model output in a similar way to GPA. However, the main difference between these two methods is that PARAFAC2 derives this rotation matrix in the latent space while GPA does it in the 'maximum' space, i.e. $\max(p_i)$.

Divergences are apparent between the scaling transformations applied in the five analysis techniques described above and a classification is proposed below to highlight three strategies available to model individual differences in scaling. The first group includes techniques applying an isotropic scaling to the individual configurations prior to the analysis. The aim of this pre-processing step is to balance the contribution of the assessors in the analysis. PCA and MFA belong to this group but they employ a different pre-scaling technique[12] In the second group of techniques the scaling procedure is built in the analysis framework and aims at weighting the contributions of the individual configurations based on a measure of agreement between

[11] W_i is referred to as a '(summed-)cross-product matrix' in Harshman (1972).

[12] It seems unclear from the literature which of these two pre-scaling techniques is preferable. The view of Kunert and Qannari (1999) is that the 'first-eigenvalue' scaling approach of MFA favors configurations with a high dimensionality but the view of Morand and Pagès (2006) is that the 'unit sum-of-squares' scaling approach favors configurations with a low dimensionality.

assessors. Such scales can be obtained by a direct computation as in STATIS or by an iterative process as in GPA. Note that the isotropic scaling applied in GPA has the role of a true 'consensus' weight only when a pre-scaling step has been applied as discussed in the previous section (Arnold, 1992). The third group of techniques follows also the idea of an individual weighting based on agreement but the scaling factors are derived for each latent dimension separately. Examples of this approach include PARAFAC2 in which individual weights are calculated during the iterative model fitting and the individual difference MDS technique (INDSCAL, Carroll and Chang, 1970) not covered in the present review but also applicable to this type of data.

The selection of a multivariate data analysis technique does not usually have a dramatic impact on the group-average pattern derived from an IV data set but the mathematical transformations applied to each individual configuration in these different methods can bring useful and complementary information about the way assessors use the sensory descriptors. Such techniques should therefore be viewed as a tool to model and assess variations between individual sensory profiles. This issue has been researched in the field of sensometrics for example by Hanafi *et al.* (2004) and Bro *et al.* (2008) with the aim of building a hierarchy of models that would allow to assess the nature and extent of various types of individual differences with a formally defined hypothesis testing framework, i.e., a multivariate analog to what is offered by the assessor model proposed by Brockhoff (2003) in the univariate domain.

7.4 Validity issues with individual vocabulary methods

In Chapter 4, the issue of sensory profiling data validity has been discussed in the context of CV methods. The same principles apply to IV methods since this sensory analysis approach is also intended to be an objective measuring instrument, that is, it employs a panel of assessors from which reliable judgments are expected. However, some of the concepts and methods presented earlier are not applicable to IV data because of the idiosyncratic nature of attributes employed in this type of experiment. The three validity criteria of **repeatability**, **agreement** and **discrimination** defined in Section 4.3 can still be considered for IV panel performance measurement but some differences need to be taken into account when translating these criteria to the univariate and multivariate domains.

Considering a set of repeated scores given by an assessor for one attribute, individual measures of reliability and discrimination can be derived irrespective of the consensus or individual nature of a sensory profile but measuring the agreement between assessors on an attribute basis only makes sense in the former case. The univariate measurement of agreement is not possible in the latter case because individual attributes are not commensurable. When dealing with an IV data set, the comparison of assessor scores and more generally the concept of panel average can only be apprehended in the multivariate domain. This means that the measures of panel repeatability, panel discrimination and agreement between assessors need to be applied on the basis of full (or partial) sensory profiles or, alternatively, after some data trans-

formation has been applied in the multivariate domain, e.g. with GPA. The idea of individual and group-average sensory configurations is therefore instrumental in the definition of validity measures for IV data and multivariate analysis methods play a major role in panel performance assessment with this type of sensory data.

In this section, several useful data analysis techniques for the univariate and multivariate assessment of IV panel performance are presented.

7.4.1 Univariate approaches to individual vocabulary panel performance evaluation

As noted in Chapter 4, assessor repeatability is the preliminary requirement for a valid sensory profile because of its direct impact on individual discrimination. It is therefore important to assess the ability of an assessor to score the objects under study in a consistent way for each of the attributes in his or her individual vocabulary, which obviously requires that replicated scores are available. Several techniques to measure assessor repeatability were presented in Chapter 4 but they are not all applicable to IV data. For example, the graphical methods proposed by Rossi (2001) or Bi (2003) are not suitable in the present case because they rely on panel mean scores for a given attribute. It is possible however to assess both individual repeatability and discrimination from a one-way ANOVA model applied to the repeated scores of an assessor for a given attribute as described by Næs and Solheim (1991). The former measure can be derived from the residual variance of the model, i.e. the mean squares of error (MSE) while the latter measure relates to the 'treatment' F-ratio or its associated p-value. It should be noted though that these two measures are not independent in this context because the F-ratio is computed from the MSE in a one-way ANOVA model. Tomic *et al.* (2007) presented several visualization methods to summarize this information with especially the p∗MSE plot well suited for a rapid detection of problematic individual attributes.

7.4.2 Multivariate approaches to individual vocabulary panel performance evaluation

The reliability and discrimination criteria discussed above provide a univariate measure of the individual performance of the assessors but the other aspects of panel performance covering the global quality of an individual sensory profile, the agreement between assessors and the discrimination at a panel level have to be tackled from a multivariate perspective. In practice, an individual sensory configuration is defined as a matrix X_i whose variables represent the individual attributes of the assessor i and are usually centered. This multivariate data representation has already been utilized for the description of the GPA procedure earlier in this chapter and it can be exploited to assess the level of performance of either a given assessor or the entire panel. Two approaches are considered below for illustration purpose with on the one hand a direct method of numerical estimation based on the individual configurations and on the other hand a technique of display in the latent domain resulting from a multivariate analysis. Note that these methods are presented in detail in this

section because they are not so commonly encountered in the literature and they will be exploited in Chapter 9.

Evaluation based on raw sensory configurations

An intuitive approach to assess the performance of an IV panel consists in obtaining numerical estimates directly from the individual configurations that can then be easily tabulated for inspection. Schlich (1996) described how such measures can be derived from the association matrix, which has already been presented in the brief introduction to STATIS given in Section 7.3.2. This matrix defined as $W_i = X_i X_i'$ characterizes entirely the relationships between objects and can be used as a basis to estimate two criteria which are respectively the dimensionality of an individual object space and the level of similarity between sensory profiles obtained from one or several assessors.

The former criterion can be considered as the measure of 'complexity' of an individual object space and can be estimated analytically by the β coefficient proposed by Schlich (1996), which is computed from the association matrix as follows:

$$\beta_i = \frac{trace(W_i))^2}{trace(W_i^2)} \tag{7.1}$$

This dimensionality coefficient can vary from 1 to $P_i = min(n-1, p_i)$, where n is the number of evaluated samples and p_i is the vocabulary size of the assessor i.

The latter criterion is based on the RV coefficient (Robert and Escoufier, 1976) and associated normalized RV coefficient (Schlich, 1996). RV is a generalized correlation coefficient defined as follows:

$$RV(W_i, W_j) = \frac{trace(W_i W_j)}{\sqrt{trace(W_i^2).trace(W_j^2)}} \tag{7.2}$$

This measure ranges from 0 to 1 and reflects in this case the similarity between two individual configurations. It can therefore be applied either to the replicated sensory profiles of a given assessor for the assessment of individual repeatability or to the sensory profiles of different assessors for the assessment of agreement. In the latter case, a matrix of RV coefficients is obtained and a measure of agreement between assessors can be derived in a similar way to the STATIS method (Schlich, 1996).

Evaluation based on a latent multivariate representation of sensory configurations

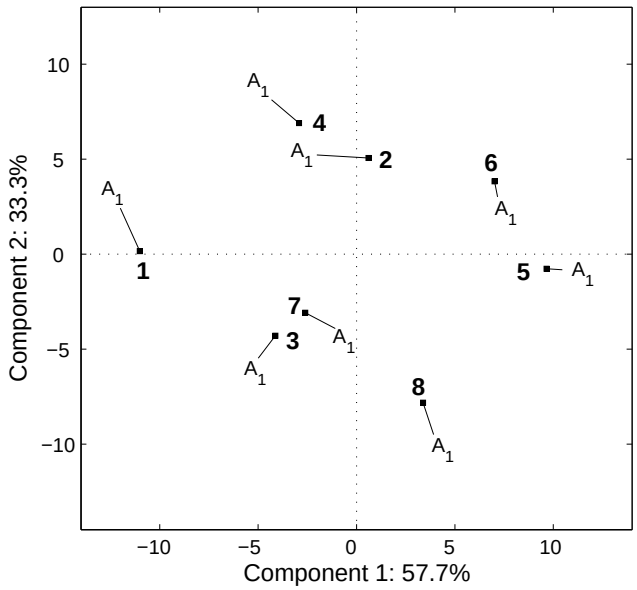
The assessment of panel performance involving a latent multivariate representation of sensory configurations offers a complementary approach to the direct numerical estimation based on the association matrix. The different multivariate analysis methods described in Section 7.3.2 of this chapter share the similar aim of representing a set of objects with respect to a small number of latent variables (principal dimensions) describing best the variation in the data under the assumptions of the analysis technique considered. When displaying the objects in this space of reduced dimensionality, a

common sensory map equivalent to a panel average is obtained. Additionally, individual configurations can be displayed in this common space to compare the sensory map of each assessor with the panel-average map. The technique to superimpose individual configurations depends on the selected multivariate analysis method but it is based on the similar idea of projection on the common subspace.

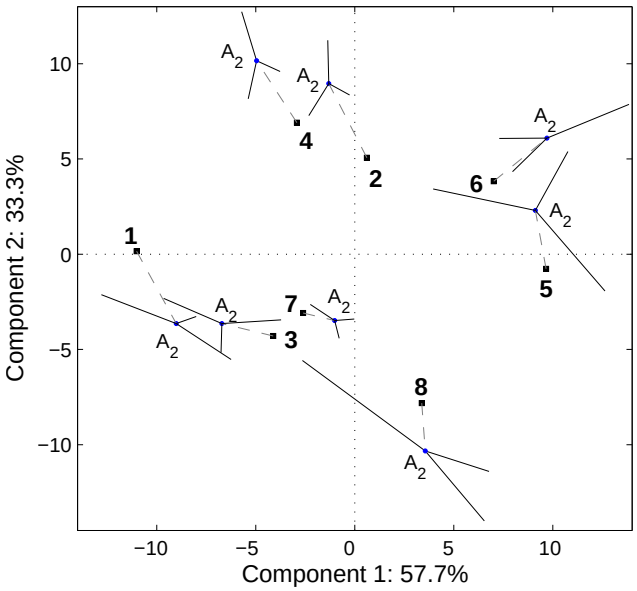
To illustrate how this principle can be exploited for multivariate panel performance evaluation, an example of IV data set is considered in which a panel of assessors evaluated a set of eight stimuli in three replicates. Firstly, a common sensory map representing the panel average is derived for example by GPA/PCA, unfold-PCA or MFA. This map is illustrated by the squares with the indices 1 to 8 in the two plots of Figure 7.3. Secondly, individual sensory configurations are projected on this common space to assess individual variability around the panel-average sensory map. Figure 7.3(a) illustrates the superimposition on the common space of the individual map (averaged over replicates) for the assessor A_1. The distance between the two sets of points highlighted by the solid lines gives an indication of the overall agreement between this assessor and the panel. Figure 7.3(b) illustrates the three replicated sensory profiles of the assessor A_2 and their associated mean. The three solid lines around each object in this plot give an indication of the multivariate repeatability of this assessor which is independent of the disagreement of this assessor shown by the dashed lines in the same plot. Similarly, the level of repeatability of the entire panel can be assessed by displaying the average across assessors for each of the three replicates separately. It should be highlighted that the distances relating to agreement and (individual) repeatability can be compared in this analysis framework because they relate to the same latent space but this would not be possible if a separate multivariate analysis was applied to the replicated profiles of a given assessor. Thirdly, a measure of discrimination can also be considered within this multivariate framework. For example, a bootstrapping method similar to the one introduced in Chapter 4 (Husson *et al.*, 2005) can be applied to define an ellipse of confidence around each object of the panel-average sensory map. This visualization technique gives a way to assess the level of discrimination between objects at a panel level.

7.5 Conclusion

In this chapter, the principles of the IV approach for verbal descriptive analysis were reviewed. Based on a literature review of the techniques frequently applied for this type of sensory analysis, three components common to IV techniques were highlighted, that is, 1) an IV development employing a more or less structured elicitation method depending on the sensory expertise of the assessors, 2) an individual attribute rating phase and 3) an interpretation of the resulting sensory profiles by the experimenter or data analyst. It was also noted that the challenge of the IV approach lies in the last step of this procedure and for this reason the aspect of IV data analysis was developed from both a qualitative and a quantitative perspective. The emphasis on multivariate data analysis became apparent from this review and the application of such analysis methods to IV panel performance assessment was also illustrated.



(a) Visual illustration of the level of agreement with the panel for the assessor A_1.



(b) Visual illustration of the repeatability of the assessor A_2.

Figure 7.3: Visual representation of a common sensory map of eight objects (indices 1 to 8) in a two-dimensional latent space with selected superimposed individual maps. The upper and lower graphs illustrate respectively the level of agreement and repeatability of two assessors.

Chapter 8

Development of a procedure for rapid individual verbal elicitation

8.1 Introduction

One of the aims of this thesis work was to explore methods for rapid sensory evaluation of audio applications. Verbal descriptive analysis using the individual elicitation approach appeared as an interesting option to consider in the continuity of the consensus vocabulary development work presented in Chapter 5 and the comprehensive literature presented in the previous chapter formed an excellent basis for the development of a suitable methodology for this type of experiment. A procedure entitled 'individual vocabulary profiling' (IVP) developed during this thesis work is presented and discussed in this chapter.

8.2 Individual Vocabulary Profiling (IVP)

8.2.1 Overview of the IVP procedure

IVP is a rapid descriptive analysis method using an individual vocabulary development approach, that is, a verbal elicitation procedure in which each subject develop his or her own set of sensory descriptors. It combines features from Free-Choice Profiling, the Repertory Grid Method and Flash Profile reviewed in Chapter 4 and can be characterized as a relatively efficient sensory profiling procedure tailored for sensory testing with inexperienced assessors[1]. The core of the proposed methodology lies in the use of a comparative evaluation technique following the approach of Flash Profile (Dairou and Sieffermann, 2002; Delarue and Sieffermann, 2004) with the rationale that using a direct comparison of the stimuli under study facilitates the elicitation process and improves the sensory discrimination. In addition, an effort is put during the individual elicitation process in gathering relevant semantic information about the vocabulary created by the assessors in the form of attribute definitions. This aspect

[1] In this thesis, the term 'inexperienced' refers to an assessor who have not been trained for the specific task of verbal descriptive sensory analysis. Following the classification of sensory assessors presented in Chapter 2, such test subjects might correspond to either *naïve* or *initiated* assessors.

has not been emphasized in the literature but it appears to be of great importance for the semantic interpretation of results. Finally, a component of the RGT methodology (Kelly, 1955) is also included in the IVP procedure as this is claimed to help assessors with no or little experience in descriptive sensory analysis (McEwan *et al.*, 1989; Piggott and Watson, 1992).

During the process of this thesis work, the author designed three sensory experiments based on an individual elicitation approach. Each of these studies contributed to the elaboration and the refinement of the IVP procedure. The first study concerns the evaluation of spatial enhancement systems for sound reproduction over headphones (Lorho, 2005b) which will be reported thoroughly in the next chapter. This study was originally designed to assess the feasibility of IVP for audio applications and the positive results of the experiment encouraged the author to explore further the methodology with two additional studies contributing to the optimization and automation of the proposed IVP procedure. One of these experiments was performed on a multimodal stimulus set relating to mobile phone slider mechanisms (Johnson, 2006), which aimed at demonstrating the applicability of the method to a different field of perceptual evaluation involving tactile and auditory aspects. The third study was designed to evaluate a set of mobile phone loudspeakers and has been reported in a publication by Lorho (2007).

An overview of the IVP procedure is presented next together with some details regarding the semi-automated system for IVP application implemented in this work.

8.2.2 Descriptive sensory analysis procedure

The whole IVP procedure can generally be completed in about three to five hours and is divided into four separate steps, as illustrated in the block diagram presented in Figure 8.1. The process starts with a phase of familiarization to the set of stimuli under study. The two next steps cover the individual process in which each assessor performs a direct verbal description of the perceived stimulus differences. In the first step of this elicitation process, a list of words[2] is generated by each assessor in a relatively free form. A second elicitation step follows during which a set of more formal attribute scales is created, i.e., sensory descriptors with a quantitative intensity scale and associated word anchors. Finally, a training phase can be considered, allowing assessors to get more familiar with the attribute scales they developed and possibly refine their vocabulary. These successive steps are described in detail below.

Selection of assessors

Prior to the individual vocabulary development, the selection of appropriate assessors for the task has to be considered. Several alternatives are possible for this phase. The first approach consists in applying a generic assessor selection procedure based on e.g. discrimination skills, reliability in grading and general verbal fluency (see Isherwood *et al.*, 2003, for an application). The selection of assessors based on general

[2]The use of the term 'word' indicates that no structured sensory concept has been created yet at this stage.

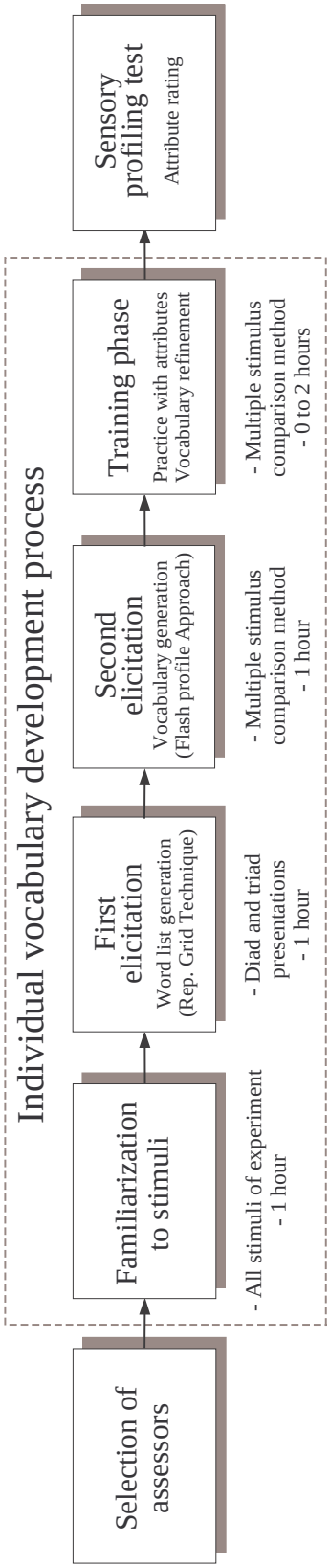


Figure 8.1: Steps of the individual vocabulary development procedure employed in this experiment. A total of 3 to 5 hours is needed for this process depending on the inclusion of a training phase or not.

discrimination skills is effective for the creation of a large pool of listeners but it might not ensure the best assessor performance for the specific set of stimuli under study. Therefore, a more focused approach consists in selecting assessors based on their ability to discriminate the stimuli under study. The author applied this method for the audio descriptive analysis experiment reported in the next chapter (see also Lorho, 2005b) to screen the ten best listeners out of twenty based on a triangle discrimination test methodology. A third alternative is to select a larger number of assessors on the loose basis of availability and interest and apply a post-screening procedure during the experiment based on reliability measures. This approach is not recommended for the selection of a consensus panel but it is conceivable for short descriptive analysis projects using the IV approach.

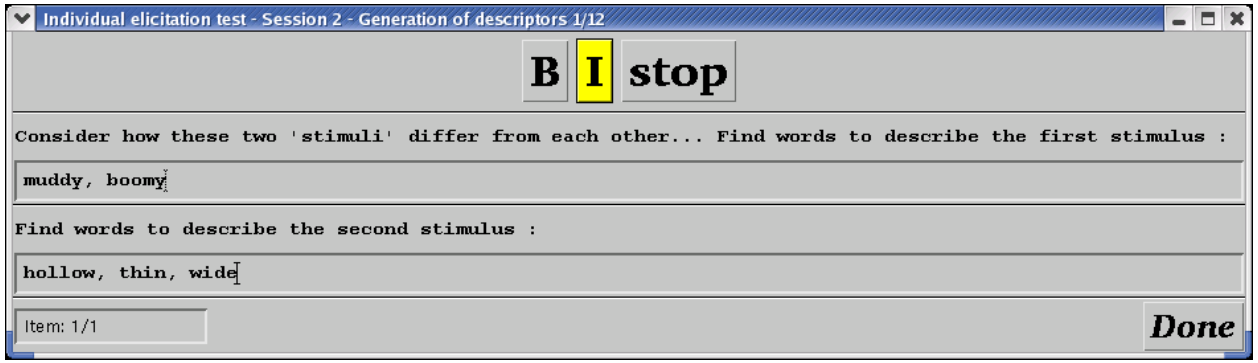
Stimulus familiarization phase

The first step of the individual vocabulary development process ensures that the assessors become familiar with the stimuli under study before the elicitation phase. Listeners can be exposed to the set of stimuli through different methods. A listening test of similarity rating can be performed to familiarize the assessors with perceptual differences between the stimuli (see e.g. Mattila (2001b) or Martens and Zacharov (2003) for an application of this method). An introduction to perceptual differences might also be indirectly given through a listening test of hedonic nature. This approach gives the assessors a chance to grade their like/dislike for the stimuli under study before they switch to a more analytical mode of thinking during the individual vocabulary development process. Alternatively, an efficient approach consists in using a discrimination test of the stimuli under study as a familiarization task, which can also serve as a pre-screening of the assessors.

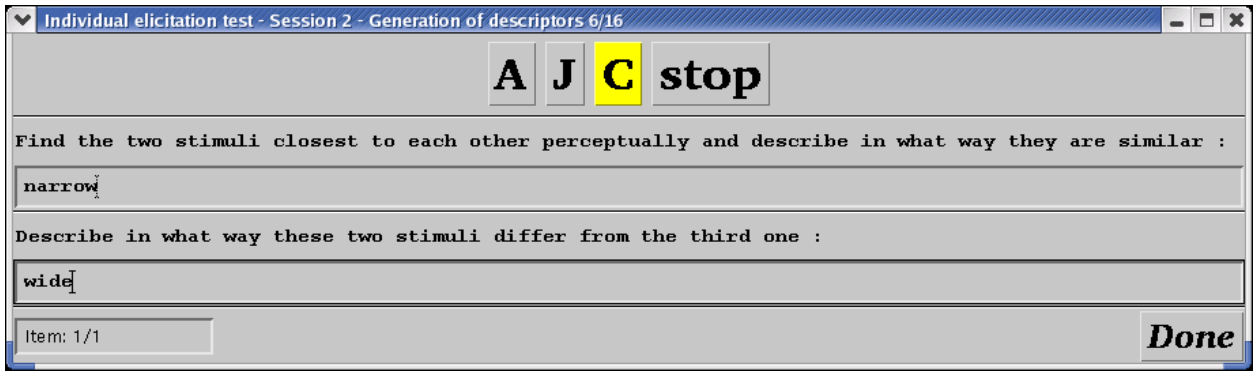
First elicitation phase

The aim of the first verbal elicitation phase is to develop an initial list of words describing differences between the stimuli under study. As the idea of this task might sound very abstract to inexperienced assessors, the gradual structure of the RGT methodology is exploited during this phase. First, a pair of stimuli is presented to the assessor who is asked to find words to describe any perceived difference between these items, as illustrated in Figure 8.2(a). The presentation of a set of ten to twenty pairs including all stimuli at least once is usually sufficient to generate a comprehensive list of words in a relatively free form. The experimenter can review this list at the end of the task to ensure that the assessor created descriptive, discriminant and non-hedonic terms. A second set of elicitation windows is then presented but now the diad presentation is replaced by the original triad method of the RGT, as illustrated in Figure 8.2(b).

The assessor is asked to focus on a single perceptual aspect differentiating any two of the stimuli from the third one, and to find two words describing respectively the perceived similarity and difference. By applying this method on a limited set of triads, e.g. ten, a more systematic description of the perceptual characteristics can be achieved. During the task, the previous list of words is available to the assessor who can either reuse some words or create new ones. The aim of these two steps is to



(a) Screenshot of the GuineaPig 3 user interface window employed for the presentation of diads in the first elicitation phase.



(b) Screenshot of the GuineaPig 3 user interface window employed for the presentation of triads in the first elicitation phase.

Figure 8.2: Illustration of the elicitation method employed in the first phase of the vocabulary development process: a series of diads is presented first (a) followed by a series of triads (b).

familiarize the assessors with the range of perceptual differences between the stimuli under study and assist them in the creation of an initial list of descriptive terms. However, these words are not utilized directly in the attribute rating phase, they only serve as a basis for the generation of a more formal set of attribute scales during the second elicitation phase.

Second elicitation phase

The second phase develops the elicitation process further with the creation of a well defined individual descriptive vocabulary. A comparative presentation of all stimuli is employed at this stage, as found in Flash Profile, but in the present case assessors can rely on the list of words they generated in the previous phase. Also, to compensate for the inexperience of the assessors in descriptive sensory analysis, an intuitive approach is adopted to introduce the concept of attribute scale. The rating method employed in this procedure is the unstructured line scale commonly used in quantitative sensory analysis methods such as QDA (Stone *et al.*, 1974). Assessors are instructed to create attribute scales to quantify the intensity of a characteristic they perceive in the stimulus set. In practice (see Figure 8.3), subjects are asked to give a name to the perceptual aspect, to define end-points for the scale, i.e., find names describing respectively a low and high quantity of this attribute, and finally to give a definition for the elicited attribute.

Following the generation of an attribute, the assessor can test the usability of the associated scale on a small set of stimuli, as illustrated in Figure 8.4. Through this procedure, the assessor can understand quickly the principle of attribute intensity scaling and is able to build a preliminary vocabulary in a relatively short time. Also, the experimenter gets relevant information about each attribute from the associated definition. At the end of this phase, a short discussion to review the vocabulary with the assessor is usually beneficial. The experimenter needs to control that all attributes are descriptive, discriminant and non-hedonic and can also make comments on the validity of the intensity scales. This approach of hedonic term screening by the experimenter has been reported in the literature by, e.g. Piggott and Watson (1992).

Attribute training phase

When the more formal set of attribute scales has been developed, the option of including or not a phase of training with the scales depends on the application. To facilitate the attribute rating task at this stage of the experiment, an attribute test with a stimulus subset can be considered, especially when the comparative evaluation includes a large number of stimuli, e.g. ten or more. As a last feature of this vocabulary development process, a review of all individual vocabularies is performed by each assessor at the beginning of the training phase or before the final sensory profiling test if the training phase is skipped. This technique, also applied in Flash profile, has often the positive effect of stimulating the description of additional aspects or can help assessors to clarify the definition of some attributes. In terms of practical implementation, a similar user interface method to the one illustrated in Figure 8.3 is employed for this training phase, which gives the assessors a chance to update the name, end-points and definition of their attributes.

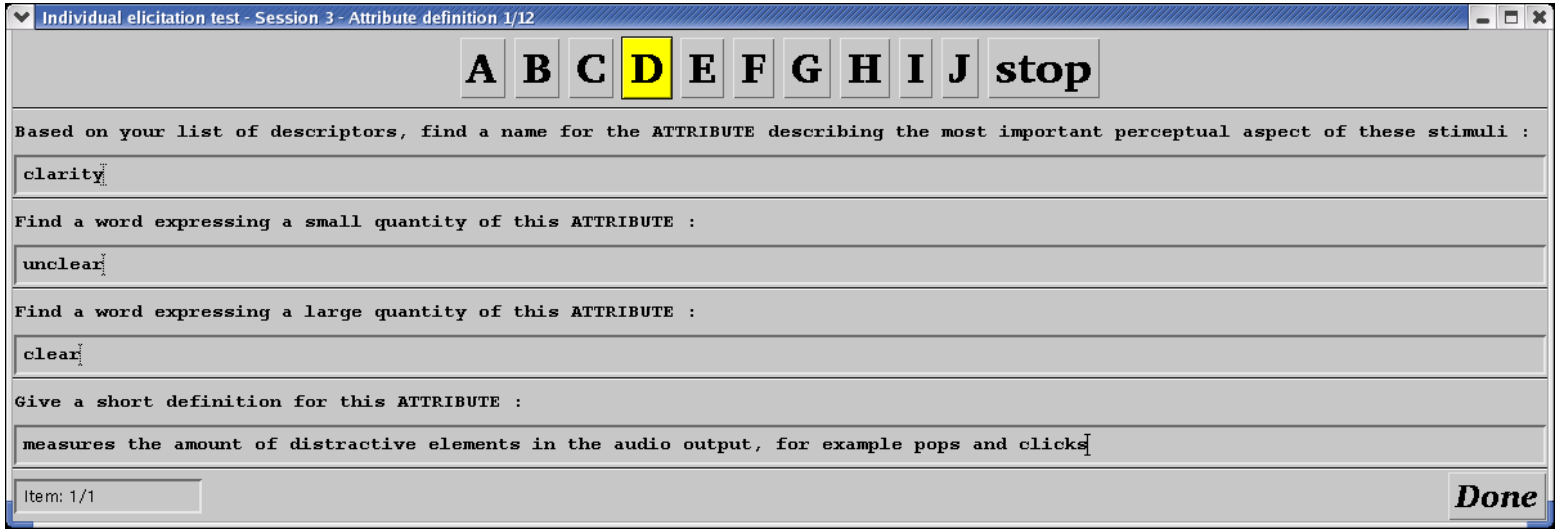


Figure 8.3: Screenshot of the GuineaPig 3 user interface window employed for the creation of an attribute scale during the second phase of the vocabulary development process. For each new perceptual aspect considered, the assessor is asked to give an attribute name and associated low anchor, high anchor and definition.

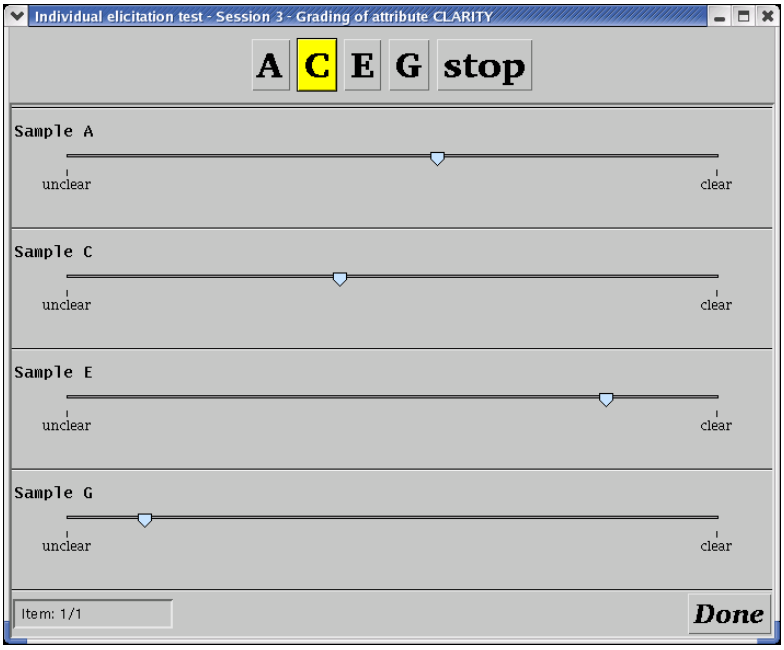


Figure 8.4: Screenshot of the GuineaPig 3 user interface window employed for the testing of an attribute scale created during the second phase of the vocabulary development process. Following the definition of an attribute illustrated in Figure 8.3, the assessor is required to grade a subset of sound stimuli to test the usability of this new attribute.

Sensory profiling test

Following this individual vocabulary development, each assessor performs the main attribute rating test with his or her final set of individual sensory descriptors. A user interface window similar to the one presented in Figure 8.4 is employed for this task but the evaluation is made with the full set of stimuli. Several sessions might be needed when more that one series of sound stimuli is included in the study. This usually occurs when a set of audio systems is tested for several audio samples (e.g. music clips) as illustrated in the study of spatial enhancement systems to be presented in the next chapter.

8.2.3 A semi-automated system for IVP test design, administration and analysis

The administration of this descriptive sensory analysis procedure is handled with the GuineaPig3[3] listening test system. The first descriptive analysis experiment of this research work (Lorho, 2005b) showed that the work of the experimenter with such an individual test procedure is proportional to the number of assessors. In practice, a significant amount of time is needed to prepare graphical user interfaces with individualized attribute scales and this work can only be made off-line. For the

[3]GuineaPig3 is an audio-visual test system developed under Linux and sharing the same principle and software structure as the GuineaPig2 listening test system (Hynninen and Zacharov, 1999).

two following IVP experiments, an automation of the test procedure was considered at three different levels.

Firstly, an automation of the GuineaPig3 user interface presentation was implemented to allow each IVP test session to be performed by any assessor without interaction with the test administrator. In practice, the successive presentation of user interface windows is controlled automatically and the text input of the assessor is stored and reused by the system when necessary. For example, the attribute name and end-points are automatically transferred from the attribute definition window to the attribute usage window (Figures 8.3 and 8.4).

Secondly, a Matlab routine was developed to generate a full IVP test structure based on few inputs such as the number of assessors, the number of stimuli, the number of elicitation windows for each step, etc. This routine creates all the GuineaPig3 scripts needed for an experiment and builds a set of command line scripts to handle three tasks: the ordered presentation of the test windows, the transfer of information from one window the next one, and the extraction of individual test results for the different sessions.

Thirdly, a Matlab routine to handle and analyze attribute rating data was also developed, which improved the speed of the IVP test result processing and interpretation.

As a result, the automated presentation method was judged positively by the assessors who participated to the most recent study using this procedure and the automated handling of the IVP process decreased considerably the time needed by the test administrator to run the descriptive analysis experiment.

8.3 Discussion

Three experiments supported the development of the IVP technique presented in the previous section and the most recent of them, that is, the perceptual study of mobile multimedia loudspeakers reported in Lorho (2007), illustrates best the potential of this descriptive sensory analysis approach. The sixteen naïve assessors who participated to this experiment were able to build an individual vocabulary of moderate complexity in three hours through the application of the systematic procedure described above. The identification of the important perceptual aspects of sound reproduced over mobile multimedia loudspeaker systems was also facilitated by the sensory descriptor definitions provided by the assessors and a good level of confidence in this interpretation was ensured by the combined analysis of the different vocabularies. Moreover, a good agreement was found with existing literature on the topic, especially the extensive set of perceptual studies on loudspeakers, headphones and hearing aids published by Gabrielsson (1979a) and Gabrielsson and Sjögren (1979).

The sensory profiling approach presented above relates closely to the methodology found in Flash Profile with regard to the use of a comparative evaluation technique but it can be said to depart from the original idea of FP toward the implementation speed. Indeed, while FP aims a minimizing the duration of the sensory profiling task to a strict minimum, that is, two sessions or even a single session, the present method

allows assessors to spend slightly more time on the development of their own vocabulary, which was found to be better suited for inexperienced assessors in the present work. In addition, issues regarding the semantic interpretation of results when applying FP, as reported by Delarue and Sieffermann (2004), appear to be reduced when a formal definition of sensory descriptors is obtained from assessors and when the number of assessors employed in the experiment is large enough. This latter aspect was apparent from the present series of three studies in which the semantic interpretation of results was found to be more robust for the study by Lorho (2007) using sixteen assessors than for the studies by Lorho (2005b) and Johnson (2006) using respectively ten and thirteen assessors.

Several benefits of the IVP technique should also be mentioned at this point. From a practical point of view, the semi-automated IVP system described above can be said to be efficient as it offers the possibility to run an unsupervised sensory profiling experiment with a large panel of inexperienced assessors, e.g. naive consumers, in a relatively short time. From the methodology point of view, the modular structure of the procedure offers offers flexibility in terms of of application of the IVP technique. For example, the procedure can be easily adapted for sensory testing with experienced assessors by removing the 'first elicitation' module (see Figure 8.1) which is only beneficial to inexperienced assessors. As a matter of fact, the individual vocabulary development process employed in the study by Lorho (2005b) was intended for selected assessors and did not include the 'first elicitation' phase (see Chapter 9 for more details). Instead, a free attribute elicitation was employed as found for example in the Flash Profile experiment reported by Delarue and Sieffermann (2004).

It should be noted finally that the comparative evaluation method employed in the IVP technique is better suited for small-size experiments, i.e., studies requiring the comparison of less than, e.g., fifteen stimuli. In practice, the simultaneous evaluation of multiple stimuli loses its efficiency for large sample size and might become too challenging cognitively, especially for inexperienced sensory assessors.

8.4 Conclusion

A rapid descriptive analysis method using the individual elicitation approach was introduced. This procedure referred to as Individual Vocabulary Profiling (IVP) combines the diad and triad comparison method found in RGT to facilitate the elicitation process for inexperienced assessors and the comparative evaluation technique of FP to improve sensory discrimination. A special attention is given in IVP to the definition of individual attributes selected by the assessors during the vocabulary development phase as this information is considered critical for the subsequent experimental step of semantic interpretation of results. In addition, IVP appears to be well suited for the unsupervised administration of sensory tests and its modular structure offers some flexibility for the design of experiments tailored for different types of sensory panels.

Chapter 9

Application of Individual Vocabulary Profiling to spatial enhancement systems - Experiment and analysis of results

9.1 Introduction

In the previous chapter, 'Individual vocabulary profiling' (IVP) was introduced as a rapid descriptive analysis method using the individual elicitation approach. This procedure is applied to the study of spatial enhancement systems for headphone reproduction in the present chapter. Following an overview of the experimental design employed for this experiment in Section 2, a comprehensive analysis of the individual vocabulary data set is provided. This review of experimental results starts with a qualitative and a quantitative exploration of the individual attributes and vocabularies in Section 3 followed by an application of the generalized Procrustes analysis to the sensory profiles of each music clip in Section 4 and a multi-way data analysis of the full data set with PARAFAC2 in Section 5.

9.2 Presentation of the experiment

9.2.1 Aim of this experiment

As mentioned in the previous chapter, the IVP experiment on the spatial enhancement systems for headphone reproduction was the first application of the individual verbal elicitation approach by the author. The idea of this study was to explore alternatives to the more conventional consensus elicitation approach described in Chapter 5 and to assess the suitability of this type of sensory evaluation for audio applications.

The scope of this IVP experiment differs considerably from the consensus elicitation work reported in Chapter 5. While a comprehensive set of stimuli was selected in the former experiment to explore in a broad sense the perceptual domain of sound reproduced over headphones, the present study focuses on a set of sound samples relating

to a specific audio application. The comparative evaluation method employed in the IVP procedure requires a set of stimuli with directly comparable characteristics. The group of headphone spatial enhancement systems for mono, stereo and multichannel audio material presented in Chapter 3 (Section 3.4.3) was considered suitable for this type of experiment and the set of stimuli already employed for the consensus attribute rating test (Chapter 6) was selected for comparison purpose.

In this section, a short presentation of the experimental design employed for this study is provided, covering the sound stimulus selection, the assessor selection procedure, the IVP implementation and the overview of the resulting data.

9.2.2 Stimulus presentation, assessor selection and listening test administration

Audio material and spatial enhancement systems

The stimulus selection for this IVP experiment comprised two separate sets of audio samples already described in Chapter 6. For recall, the first subset comprises three music clips applied to eight reproduction techniques illustrating different types of spatial enhancement over headphones (i.e., the unprocessed stereo, five spatial enhancement systems, a monophonic downmix, and a mono-to-3D algorithm) and the second subset comprises one music clip and one movie sound clip applied to eight reproduction techniques relating to virtual 5.1 sound systems for stereo headphone reproduction. This gave a total number of forty sound samples covering the important perceptual aspects of this type of audio application. In the context of the individual elicitation task included in this IVP experiment, the two following remarks should be made about this full set of stimuli: 1) the perceptual differences between systems for a given audio clip appeared to be more important than the differences between audio clips and 2) the two subsets generated quite similar perceptual aspects despite the slightly different nature of the two audio applications considered.

Assessor selection

A screening procedure covering the two important aspects of discriminative and descriptive skills was employed to select the assessors of this study (see Lorho, 2005b). An audiometric test would have been difficult to run in the context of this experiment but it was assumed that assessors showing good skills in the discrimination task would be reliable for this perceptual evaluation task. The screening test was applied to twenty subjects who had not participated in an audio experiment previously. Six of these candidates were experienced in descriptive analysis and can be classified as *initiated* assessors while the other subjects were university students selected for their interest in this experiment and should therefore be considered as *naïve* assessors.

A discrimination task based on the triangle test methodology (ISO 4120, 2004) was applied with set of four stimulus pairs from this study. Each pair was chosen from the eight processed samples of a given audio clip and the selection was made to include differences varying from very clear to hardly perceptible, considering both timbral and spatial aspects of headphone sound reproduction. This increasing level of perceptual

difference was indeed confirmed by the result of the discrimination task as the first, second, third and fourth stimulus pairs were discriminated significantly by 20, 18, 14 subjects and 2 subjects respectively.

To complete this screening session, the experimenter asked the listeners to comment on the perceptual differences existing between the two samples of each stimulus pair at the end of the discrimination test. This procedure allowed to assess informally the descriptive skills of each assessor and revealed that some assessors were more confident than others in describing the spatial and timbral differences between and within stimulus pairs.

The outcome of this screening procedure can be summarized as follows. The six subjects experienced in descriptive analysis were selected because they all showed very good discrimination skills and four of the inexperienced candidates showing the best discriminative and descriptive skills were also chosen. It should be noted that the size of the panel selected for this individual vocabulary profiling experiment, i.e., ten assessors comprising three male and seven female subjects, was limited for practical reasons relating to time management.

Listening test administration

The processed stereo samples employed for this listening experiment were were stored as 16-bit PCM wave files at 44.1kHz sampling frequency and for each program item an accurate time alignment was performed to allow for direct switching between stimuli. A loudness alignment was also applied to ensure a similar inter-algorithm loudness at a comfortable listening level. A real-time implementation (Tuomi and Zacharov, 2000) of the Moore steady state loudness model Moore *et al.* (1997) was employed for the loudness alignment.

The experiment was set-up in a quiet office room and the test administration was performed using the GuineaPig3 listening test system. The reproduction chain of the stereo audio output included the following steps: digital-to-analog conversion by a Digigram sound card; balanced stereo signal amplification by a Symetrix SX304 headphone amplifier; and sound reproduction through high-quality, diffuse-field circumaural headphones (Sennheiser HD580s).

9.2.3 Implementation of the IVP procedure

The attribute elicitation methodology employed for this descriptive analysis experiment follows the principle of the IVP procedure presented in Chapter 8 but did not include the gradual structure of the RGT methodology. This approach can be justified by the fact that 60% of the assessors participating to this experiment were already experienced in sensory analysis. It should be noted however that enough time was given for the verbal elicitation phase to ensure that inexperienced assessors would cope with this unstructured elicitation process.

The individual vocabulary development process was organized in four parts covering a stimulus familiarization, a word elicitation phase, an attribute development phase and an attribute training phase. Three to four one-hour sessions were needed to complete this work, depending on the time spent by the assessors for the attribute

training. The final attribute rating experiment followed this vocabulary development process and took four to eight hours depending on the size of the individual vocabulary. More details about the attribute development process are given next.

Individual attribute elicitation

The elicitation part of this experiment was performed in two sessions of one hour each. Discussions with the experimenter were carried out in English, but assessors were encouraged to perform their elicitation in their native language, i.e. Danish for nine of the participants and English for one participant. Following a review of the attributes between the experimenter and each assessor, a translation of the final attributes to English was agreed in most cases. This language conversion was intended to facilitate the test administration work of the experimenter during the attribute rating experiment and the subsequent semantic interpretation of the results.

The first step of the individual attribute elicitation consisted in a 30-minutes introduction to sound reproduction over headphones. Fifteen pairs of audio stimuli were selected to illustrate a wide range of perceptual differences, including music, movie sound, speech, and environmental recordings with very different spatial sound characteristics. These audio samples were chosen from the headphone sound database described in Chapter 5 (see Section 5.2.1). Assessors were invited to carefully consider the differences in sound reproduction aspects within and between the sound pairs.

The second step of the procedure was completed during the same session and consisted in a preliminary individual elicitation performed in 30 minutes. Only the audio stimuli under study were considered in this task, that is, 40 stimuli divided in 5 series of 8 systems, as described above. Assessors were encouraged to come up with terms illustrating differences between the systems for a given audio item, which implies that the focus was put on finding discriminant descriptors. The eight systems presented in a window were listed from A to H and the the listener could switch freely between the systems. Word elicitation was performed on a sheet of paper and some of the assessors already created groups of perceptually related terms at the end of this session.

In the third step of the elicitation process run in a separate session, the assessors used about 30 more minutes to complete the individual vocabulary development. Following this elicitation work, the proposed vocabulary was reviewed with the experimenter. The individual attributes were described in English and word anchors were selected to describe a low and high intensity of each descriptor. Most of the assessors felt comfortable with the resulting attributes at this stage, and the last 15 minutes of this session were employed to experiment informally with the attribute scales on few audio samples. Assessors retained between 4 and 9 attributes at the end of this process. It should be noted however that only an informal definition of the attributes was included in this process, which differs slightly from the systematic attribute creation of the procedure presented in the previous chapter.

Attribute rating phase

Following this verbal elicitation part of the experiment, a training phase was considered to familiarize the assessors with their individual attribute scales. A set of user

interfaces was designed for one of the audio items (music clip #1). Figure 9.1 illustrates the user interface window employed for this attribute rating test, which was implemented using the GuineaPig3 listening test system. A random permutation was applied to associate the eight systems to the letters A to H for each attribute and the assessors were asked to perform a comparative assessment of these eight stimuli for a given attribute. The interface allowed for direct switching between the test stimuli using a 20ms exponential cross-fade with a switching latency of 200ms.

The final attribute rating test was performed after the training phase assuming that the assessors had become familiar enough with the different systems under evaluation and their individual vocabulary at this stage of the experiment. The listening test was administrated in a similar way to the attribute training session using a one-hour test session for the evaluation of one or two audio clip depending on the assessor pace and the size of his/her individual vocabulary. A repetition of two music clips from the first subset of audio samples was also included in this study to obtain a measure of assessor and panel repeatability. As a result, seven series of eight systems were assessed in a duration of three to seven sessions depending on the assessor.

9.2.4 Overview of the resulting set of data

Earlier in this chapter, it was mentioned that two different groups of audio samples were considered in the present IVP test, that is, a set of spatial enhancement systems for headphones and a set of virtual 5.1 systems for headphones. It should be noted that the assessors were not aware of this grouping and were required to apply their individual vocabulary to all the audio clips without distinction. This approach did not cause any problem to the subjects, except in the case of one individual attribute specific to the movie sound item of the virtual 5.1 study[1] and considered not appropriate by the assessor for the other audio clips. The overall suitability of the individual vocabularies for the two different stimulus sets gives an indication that spatial enhancement systems and the virtual 5.1 systems for headphones are relatively similar perceptually.

In the present chapter, only the results of the group of audio samples relating to the headphone spatial enhancement systems are presented for consistency with the consensus vocabulary data analysis presented in Chapter 6. This comprehensive data subset is considered sufficient to demonstrate the applicability of the IVP method to a set of audio stimuli and illustrate the subsequent data analysis of this type of experiment.

The outcome of the sensory evaluation made on this subset of stimuli can be described as a data structure comprising eight systems evaluated for three different music clips by 10 assessors, each of them using their own set of attributes, as illustrated in Figure 9.2. Note that the sensory evaluation of the music clips #1 and #2 was repeated by all assessors in separate sessions but this data is excluded from the present analysis.

As the same systems and music clips were employed in the experiment reported in Chapter 6 and in the present analysis, a similar coding scheme is applied, that is,

[1]This attribute is the *Sense of movement* of the assessor #10.

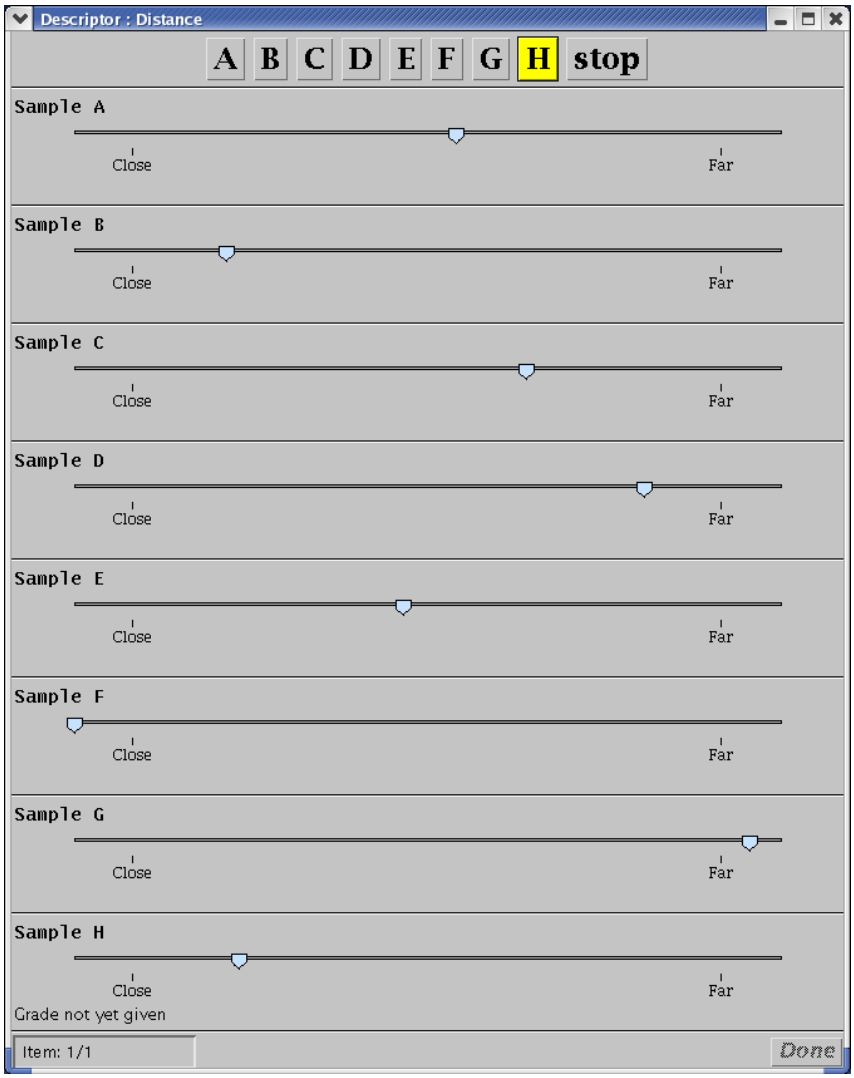


Figure 9.1: Example of GuineaPig 3 user interface employed for the attribute test.

1_S for the stereo unprocessed material, 2_M for the monophonic down-mixed version of the stereo material, 4_{Me} for the spatial enhancement algorithm of monophonic music material and 3_{Se} to 8_{Se} for the five stereo enhancement algorithms[2]. Also, the music clips are referred to as the 'music clip #1' for the *Scritti Politti* track, the 'music clip #2' for the *Madonna* track and the 'music clip #3' for the *Tuck & Patti* track.

In Figure 9.2, the sensory profile obtained from the assessor i ($i = 1, ..., 10$) for the music clip q ($q = 1, 2, 3$) is represented as a rectangle corresponding to a (two-way) matrix X_{iq} comprising 8 rows, i.e. the systems under study, and p_i columns, i.e., the number of attributes elicited by the assessor i. Note that in this chapter, X_i will often correspond to the 'centered' configuration of the assessor i for a given music clip, which means that the columns have been centered across systems by setting to zero the mean of the eight scores given for each attribute.

[2]The three perceptual flavors represented by these five systems were listed in Chapter 6 (Section 6.2.1)

The presentation of results considered in the next sections follows the analysis principles presented in Chapter 7. An overview of the individual vocabularies generated in this experiment is provided first and the results are studied qualitatively and quantitatively to identify perceptual groups from the complete set of individual attributes. A quantitative analysis of the ten individual sensory configurations obtained for each music clip follows based on the GPA technique described in the previous chapter. Finally, a global quantitative analysis is applied to the combined data comprising the three music clips with the PARAFAC2 technique also introduced earlier.

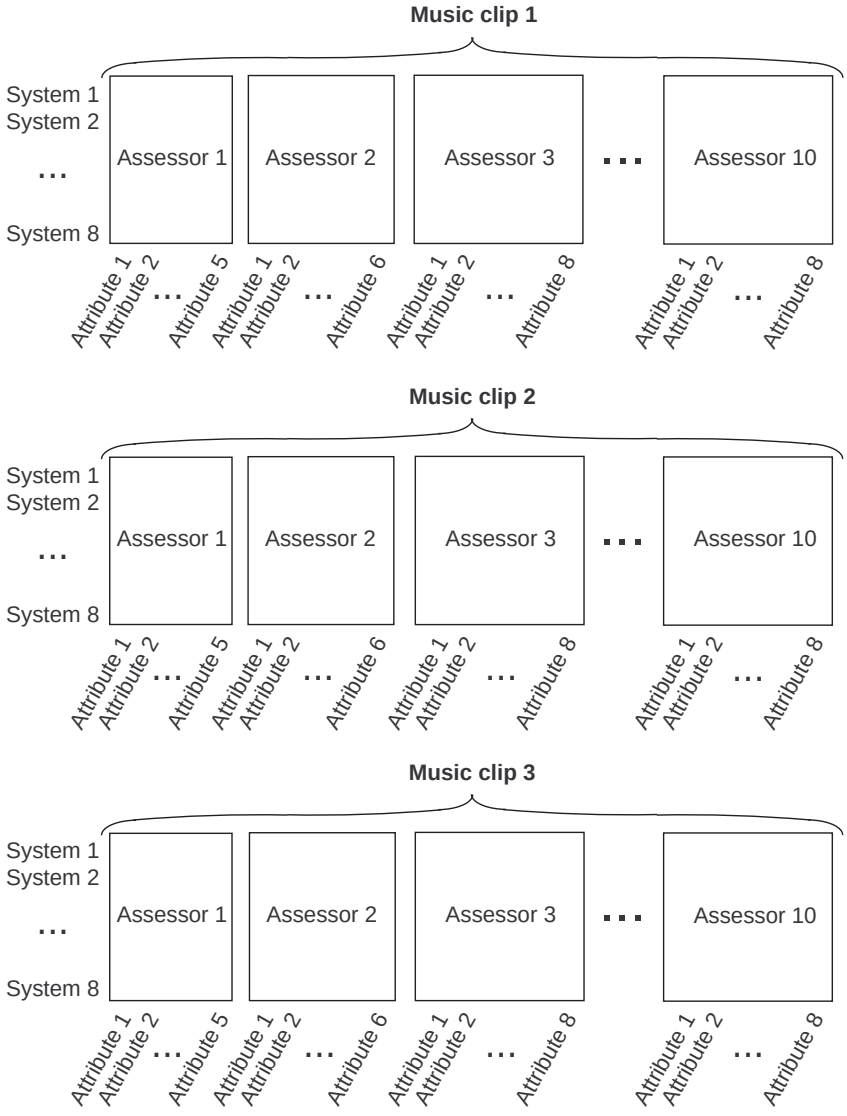


Figure 9.2: Illustration of the data set considered in this chapter which comprises a set of 10 individual sensory profiles for each music clip with an additional replication for the music clips #1 and #2. It can be noted from this data structure that each assessor employed the same individual vocabulary for the sensory evaluation of the three music clips.

9.3 Presentation and grouping of individual attributes

9.3.1 Individual vocabularies and sensory profiles

A total of 67 sensory descriptors resulted from this descriptive analysis experiment. The individual vocabulary developed by each of the 10 assessors is presented in Table 9.1 with the name and two associated end-points of each descriptor. The size of the vocabularies range from 4 to 8 attributes. Eight of the assessors elicited 6 or more descriptors while the assessor #4 generated the most compact vocabulary comprising only 4 attributes.

Table 9.1: List of individual vocabularies.

Assessor	Attribute name	Word anchors
1	Depth	No depth/A lot of depth
	Metallic	Not metallic/Very metallic
	Distance	Close/Far
	Fullness	None/A lot
	'Klang' (sound duration)	Normal/Very long
2	Bass	No bass/A lot of bass
	Power	Flat/Powerful
	Inside-Outside	Inside/Outside
	Distance	Close/Far
	Room size	Small/Large
	Reverberation	No reverb/A lot of reverb
3	Bass	No bass/A lot of bass
	Metallic	Not metallic/Very metallic
	Noise	No noise/A lot of noise
	Muffling	Not muffled/Very muffled
	Sound impression	Simple/Complex
	Distance	Near/Far
	Roominess	Narrow/Wide
	Echo	No echo/A lot of echo
4	Bass	No bass/A lot of bass
	Reverberation (Rumklang)	No reverb/A lot of reverb
	Distance	Near/Far
	Mono-Stereo-Surround	Mono/Stereo/Surround
5	Bass	No bass/A lot of bass
	Treble	No treble/A lot of treble
	Voice/sound clarity	Weak/Strong
	Voice/sound distance	Close/Far
	Room	Small/Large
	Echo	No echo/A lot of echo

Continued on next page

Table 9.1 – *continued from previous page.*

Assessor	Attribute name	Word anchors
6	Bass	No bass/A lot of bass
	Brightness	Not bright/very bright
	Muffling	Not muffled/very muffled
	Width	Mono/Stereo
	Room size	Small/Large
	Background noise	No noise/A lot of noise
	Echo	No echo/A lot of echo
7	Bass	No bass/A lot of bass
	Treble	No treble/A lot of treble
	Noise	No noise/A lot of noise
	Externalization	Inside/Outside
	Reverb	Small room/Large room
	Left-Right	Left/Centered/Right
	Sharpness	Short/Long
	Muffling	Not muffled/Very muffled
8	Bass	No bass/A lot of bass
	Treble	No treble/A lot of treble
	Clarity-Punctuation	None/Full
	Concentration of sounds	Minimum/Maximum
	Depth	Shallow/Full
	Echo	No echo/A lot of echo
	Metallic	Not metallic/Very metallic
9	Bass	No bass/A lot of bass
	Brightness	Not bright at all/very bright
	Homogeneity	Not homogeneous/ Fully homogeneous
	Sharpness	Hard-square/Soft-round
	Realism (naturalness)	Artificial/Natural
	Broadness	Compact/Broad
	Distance	Close/Far
	Movement	Passive/Active
10	Bass	No bass/A lot of bass
	Middle	No middle/A lot of middle
	Treble	No treble/A lot of treble
	Clarity	Muffled/Clear
	Noise	No noise/A lot of noise
	Reverb	Small/Large
	Width	Narrow/Wide
	Sense of movement (3D)	Absent/Present

A visual exploration of the raw data generated in this type of experiment is always a useful preliminary step to assess the relative complexity of the individual sensory profiles and the 'spider web' plotting technique presented in Chapter 5 is very appropriate for this purpose. An illustration of this plotting technique is given in Figure 7.1 for a subset of assessors and the systems 1_{S}, 2_{M}, 4_{Me} and 5_{Se} for the music clip #1.

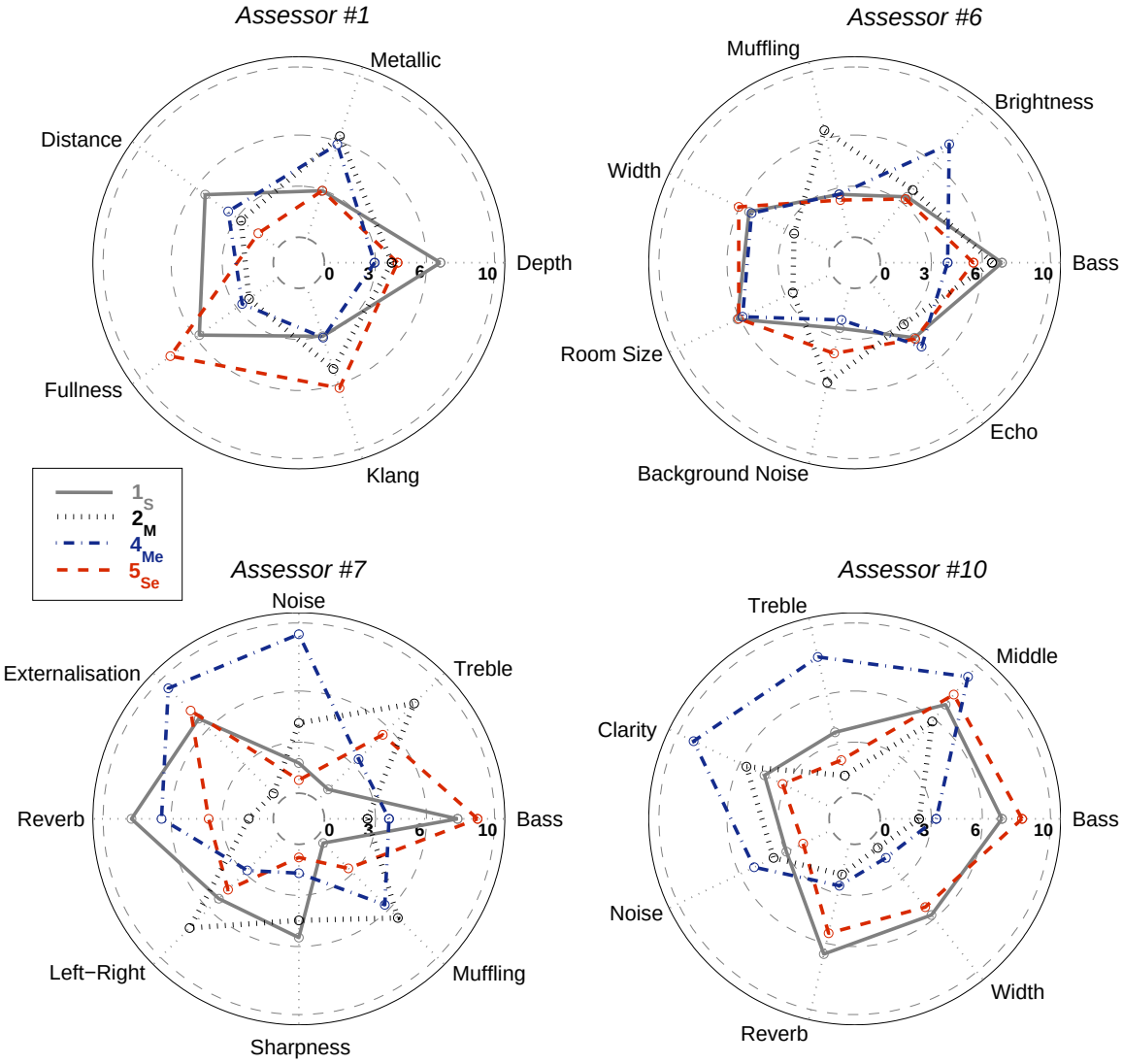


Figure 9.3: Individual spider web plot of a subset of systems and assessors.

Looking at a single spider web, one can get some indication of the way an assessor employs the scale and perceives the different samples, e.g. the assessors #1 and #6 tend to use a lower range of the 10-point scale than the assessors #7 and #10 and the assessor #7 tends to be more discriminative on this subset of systems. It can also be useful to inspect the level of similarity between scores given by different assessors for attributes with the same name. In the present case, it can be seen that the attribute *Bass* is applied in a similar way by the assessors #7 and #10 but the two other assessors show a different rating pattern for this attribute. It also appears that the

algorithm 4_{Me} has a distinctive profile with noticeable variations between assessors, e.g. assessors #7 vs. #1.

A complementary approach to the visualization of raw data consists in evaluating analytically the complexity of the individual sensory profiles. The two following methods were considered in this study. First, the β coefficient introduced in Chapter 7 was derived. This dimensionality coefficient proposed by Schlich (1996) gives a good estimation of the sensory complexity required to describe the set of samples under study. Then, a PCA was applied on each centered configuration X_i (where i is the assessor index) for a given music clip and the number of relevant principal components was estimated.

Table 9.2 gives a summary of the measures obtained for each assessor and music clip. Principal components explaining more than 10% of variation in the data were retained for the PCA dimensionality[3]. This table shows a decrease in both measures of sensory profile complexity on average from the music clip #1 to the music clip #3. The assessors #4 and #10 have respectively the lowest and highest dimensionality globally. The low score of the assessor #4 might be explained by its small vocabulary size but it should be noted that this type of correlation is not automatic as illustrated by the similar β coefficient of the assessors #9 and #4 for the music clip #2 although the size of these two vocabularies differs by a factor of two.

Table 9.2: Analysis of individual vocabulary complexity by measure of β coefficient and PCA dimensionality.

		Music clip #1		**Music clip #2**		**Music clip #3**	
Assessor	**Vocab. size**	β	**PCA dim.**[a]	β	**PCA dim.**[a]	β	**PCA dim.**[a]
1	5	3.24	3	2.41	3	1.64	2
2	6	3.44	3	3.78	4	2.39	2
3	8	3.75	3	3.87	3	2.95	3
4	4	1.92	2	1.96	3	1.66	2
5	6	2.71	3	2.32	2	1.96	2
6	7	2.83	3	2.80	3	3.16	4
7	8	3.31	3	2.89	3	2.95	3
8	7	3.24	4	3.37	3	3.01	3
9	8	2.62	3	1.79	2	2.79	3
10	7	4.02	4	3.44	4	3.27	4
Average	6.6	3.11	3.1	2.86	3.0	2.58	2.8

[a] PCA dimensionality: number of principal components explaining more than 10% of data variation.

[3]The choice of a 10% explained variance threshold is purely arbitrary. A more reliable approach would be to estimate the significant level of each principal component by a permutation technique as found in e.g. Dray (2008) but this has not been pursued here.

9.3.2 Grouping of individual attributes

The outcome of an IVP experiment is always characterized by an abundance of descriptive terms as illustrated by the above presentation of the individual vocabularies and profiles. In parallel to the assessor-based examination of these sensory profiles, comparing the attributes of different assessors can help getting a more global picture of the perceptual aspects elicited by the panel. In this section, a qualitative classification of the sensory descriptors generated in the experiment is presented first. The validity of this grouping is then tested by applying two different quantitative methods enabling the study of relationships between individual attributes based on ratings.

Qualitative grouping of attributes

A classification of the individual attributes listed in Table 9.1 was applied on the basis of a qualitative interpretation of the descriptor names and associated word anchors plus the additional comments made by the assessors during the attribute review with the experimenter. Eight perceptual aspects were identified as illustrated in Table 9.3, that is, *tone color*, *timbre*, *localization*, *room perception*, *externalization*, *broadness*, *artifacts* and *temporal aspects*. Five attributes were left out from this classification due to their ambiguous, holistic or idiosyncratic character. The attributes with the same name and word anchors have been grouped in this table, i.e., the attributes *bass*, *treble*, *echo*, *muffling*, *distance*, *room size* and *noise* elicited respectively by 9, 4, 4, 3, 3, 2 and 2 assessors.

Table 9.3: Qualitative grouping of attributes.

Group	Attribute name	Assessor	Comments
1 - Tone color	Bass	2, 3, 4, 5, 6, 7, 8, 9, 10	
	Depth	1	
	Middle	10	
	Treble	5, 7, 8, 10	
	Metallic	1	Related to treble
		3	Described as 'treble without bass'
	Brightness	9	Related to treble
2 - Timbral aspects	Muffling	3, 6, 7	
	Clarity	10	
	Voice/sound clarity	5	
	Brightness	6	
	Metallic	8	
	Sharpness	9	

Continued on next page

Table 9.3 – continued from previous page.

Group	Attribute name	Assessor	Comments
3 - Localization aspects	Distance	1, 2, 9, 4	
	Voice/sound distance	5	
	Movement	9	
	Sense of movement	10	Scale only used for the movie sound sample
	Left-Right	7	
4 - Room perception	Echo	3, 5, 6, 8	
	'Klang'	1	
	Room size	2, 6	
	Width	10	
	Room	5	
	Roominess	3	
	Reverb	10, 7	
	Reverberation	2, 4	
5 - Externalization	Inside-Outside	2	
	Externalization	7	
	Distance	3	
6 - Broadness	Fullness	1	Described as a spatial aspect
	Mono-Stereo-Surround	4	
	Width	6	
	Broadness	9	
	Concentration of sounds	8	Defined as horizontal spread and described as a broadness aspect
7 - Artifact aspects	Noise	3, 10	
		7	Described as metallic, fricative sounds, degraded sound quality...
	Background noise	6	
8 - Temporal aspects	Sharpness	7	
	Clarity-Punctuation	8	Relating to temporal distinction between sounds

Continued on next page

Table 9.3 – continued from previous page.

Group	Attribute name	Assessor	Comments
Other aspects	Realism (naturalness)	9	
	Sound impression	3	
Unsorted aspects	Power	2	
	Depth	8	Attribute related to resonance
	Homogeneity	9	

The perceptual aspects of *tone color* and *room perception* form two relatively large groups with respectively 18 and 14 individual descriptors while the groups relating to *timbre* and *localization* include 8 attributes. The four remaining groups are much smaller with less than 5 attributes and their validity can therefore be questioned. Note that the full classification should also be considered with caution as emphasized in Chapter 7. Indeed, an analysis of individual sensory descriptors by the experimenter based on qualitative aspects alone is always prone to interpretation bias. The terminology of Gaines and Shaw (1993) summarized in Figure 7.1 illustrates this aspect with the problematic scenarios of *conflict* arising when the same term is assigned to different distinctions and *correspondence* arising when different terms are assigned to the same distinction.

Attribute grouping by hierarchical clustering analysis

Hierarchical clustering is a technique commonly used for the analysis of Repertory Grid data (Shaw, 1980; Berg and Rumsey, 1999b). This classification method works by grouping entities based on their similarity or dissimilarity. In the context of sensory analysis, entities can either represent the objects under study or the individual attributes.

In the present study, the agglomerative hierarchical clustering (AHC) approach was selected with the usual Euclidean distance as a dissimilarity measure and Ward's method as an aggregation criterion. The three set of sensory profiles shown in Figure 9.2 were considered for this analysis. Prior to the AHC, each sensory profile was preprocessed separately by centering and scaling data across systems (i.e., for each attribute separately) in order to reduce differences in scale usage between assessors and music clips. A matrix of 24 objects (i.e., 8 algorithms $\times$ 3 music clips) by 66 variables (i.e. attributes) was then created and the AHC routine was applied to the variables of the matrix. It should be noted that running the cluster analysis to this combined data set gave a more stable and interpretable result than when applying separate analyses on the three sets. Also, a matrix of 24 objects by 66 variables was selected rather than a matrix of 8 objects by 198 variables (i.e. 66 attributes $\times$ 3 music clips) to facilitate the graphical interpretation of the results. AHC on objects instead of variables was also considered in this study to gather information about similarities and differences between algorithms at a global level or for each music clip separately.

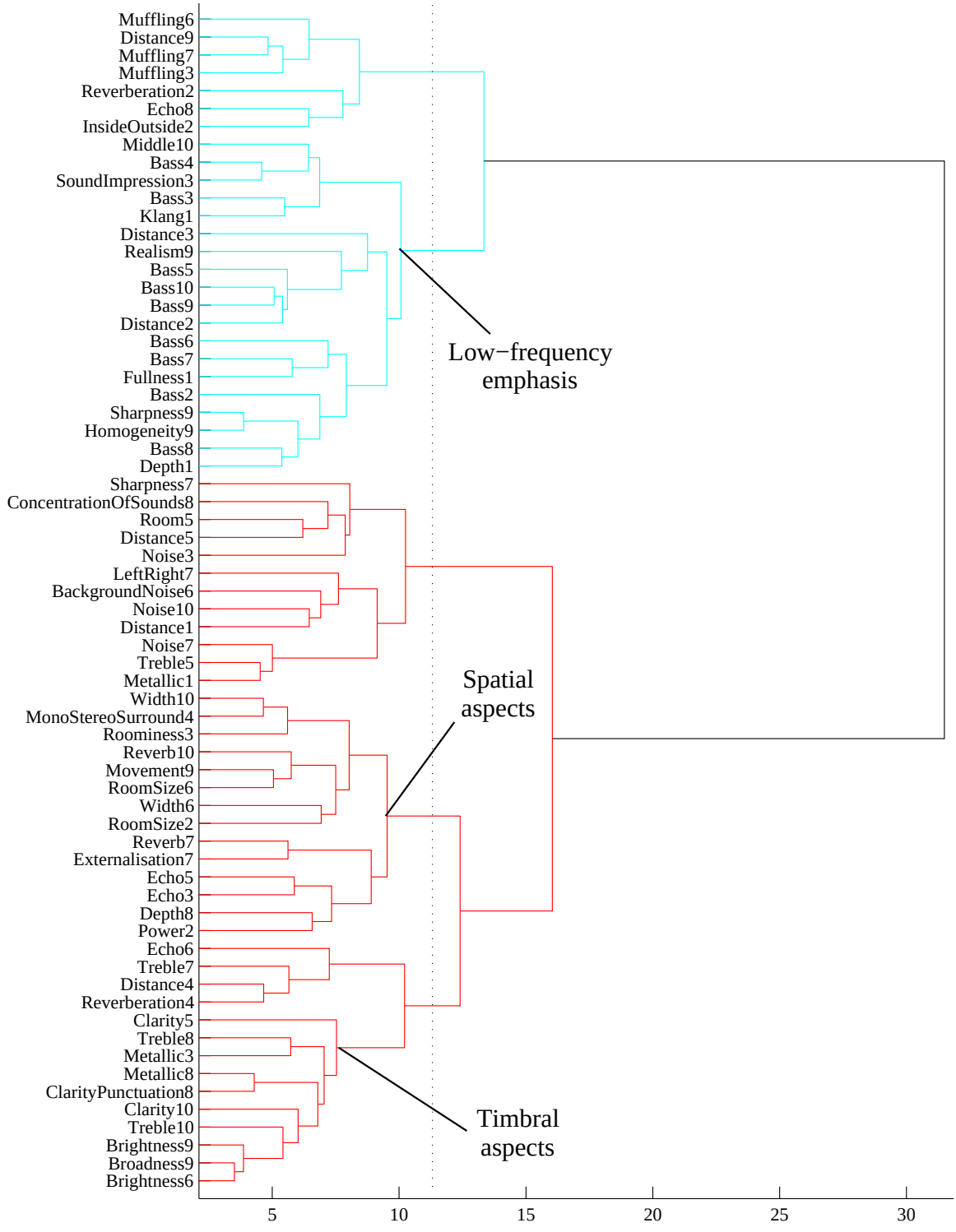


Figure 9.4: Dendrogram obtained by an agglomerative hierarchical clustering analysis of the 66 individual attributes employed in the IVP study. Five clusters with a distance measure greater than 12 (vertical dotted line) can be identified on this graph but only three of these groups of attributes can be interpreted, which relate respectively to *timbral aspects*, *spatial aspects* and *low-frequency emphasis* as shown on the figure.

Figure 9.4 shows the resulting dendrogram employed to represent the dissimilarity level between the individual attributes. Neighboring variables on the vertical axis of this dendrogram show the greatest similarity and the horizontal axis gives a quantitative measure of the distance between groups of attributes. Five clusters with a distance measure greater than 12 can be identified in this graph. An inspection of these clusters reveals that the first group (at the lower side of the dendrogram) contains 10 attributes related mostly to timbral aspects, e.g. *clarity*, *brightness*, *treble*, *metallic* and a smaller sub-group apparently concerned with spatial aspects, e.g. *echo*, *reverb*, *distance*. The closest group to this first cluster relates mainly to spatial aspects e.g. *room size* and *echo*, and contains 14 individual attributes. Moving up in the classification, a third cluster is found which includes all four occurrences of the term *noise* and a mix of spatial and timbral attributes. The fourth cluster is the largest one with 26 attributes. Based on the nine occurrences of the attribute *bass* in this group, the perceptual aspect of *low-frequency emphasis* can easily be identified. However, this cluster also includes few apparently unrelated attributes such as *distance* and *sharpness*. Finally, the fifth cluster (at the upper side of the dendrogram) is the closest to the previous group and contains attributes of different characteristics with three occurrences of the term *muffling*.

The study of this dendrogram illustrates that three perceptual characteristics can be interpreted from this sensory data set, namely **low-frequency emphasis**, **timbre** and **spatial aspects**. However, the presence of unrelated individual attributes in each of these clusters indicates that some descriptors are semantically ill-defined in the IVP data set. Also, this quantitative analysis shows a much lower level of details in comparison to the qualitative grouping presented earlier which indicates an interpretation bias in the qualitative analysis. Note however that some of the discrepancies seen between the two classifications, especially the separation between the *bass* attributes and the other attributes relating to tone color such as *treble*, are due to the way the selected AHC algorithm operates. Indeed, the polarity of the attribute scales has a large influence on this clustering process which prevents inversely correlated attributes to be grouped together. This can be considered as a limitation of selected AHC dissimilarity measure for the identification of related sensory concepts irrespective of their scaling polarity. Note however that other dissimilarity measures not investigated in this study might handle differently the polarity of attributes. The 'Focus' procedure developed specifically for the analysis of RGT data by Jankowicz and Thomas (1982) appears to model explicitly this attribute polarity aspect.

Principal component analysis of the two main perceptual groups

For the second quantitative analysis study of this section, a broader division of the individual attributes was made into two basic perceptual categories, namely timbre and spatial characteristics and a separate study of the attribute relationships was considered for each of these two categories. The first group contains the 26 attributes relating to tone color and timbre (groups #1 and #2 of the classification presented in Table 7.2) while the second group includes the 29 attributes from the group of localization (group #3) and the three groups relating to spatial aspects (group #4 to #6). A data formatting similar to the one used in the AHC analysis was applied

here and a subset of variables corresponding to the perceptual group of interest was selected from the full matrix of 24 samples by 66 centered and scaled attributes.

Principal component analysis (PCA) can be seen as a data reduction technique in which a set of observed variables is represented by a smaller set of latent variables. These new variables can be used as a basis to visualize the objects and the original variables in a space of lower dimensionality. In the present case, the main interest was to study the relationships between the individual attributes in this low-dimensional space and for this purpose the correlation loading plot approach (Martens and Martens, 2001) appeared as an intuitive visualization method. A separate PCA was applied to the two matrices defined above and the resulting attribute correlation structure was studied in each case. It should be noted that the multivariate data analyses applied to the sensory profiling data later in next sections will give a complementary view of the full set of individual attributes.

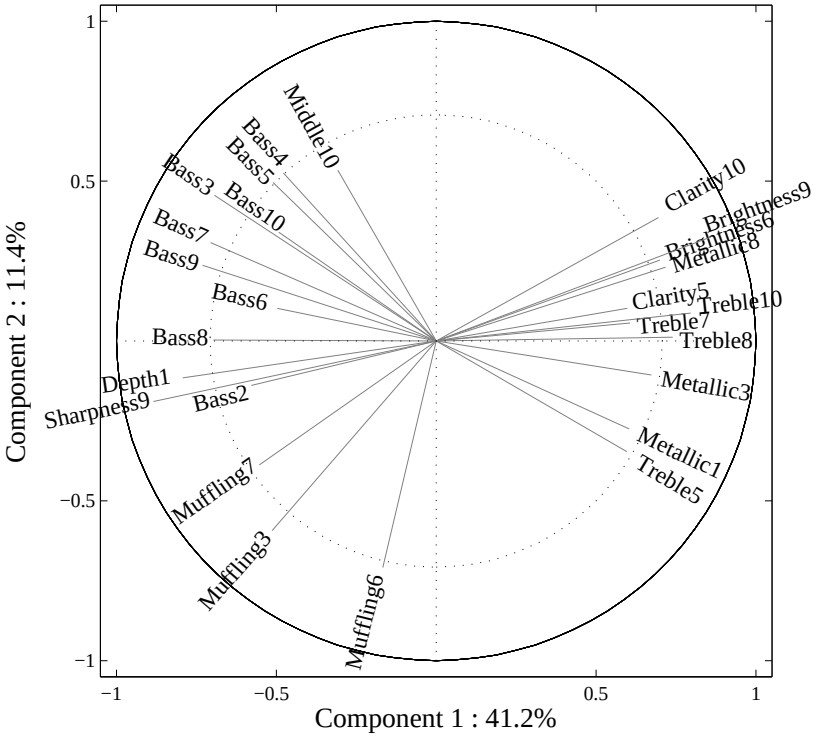
The variation in the data explained by 4-component PCA model was 41.2%, 11.4%, 6.7% and 5.9% for the category of timbre and 26.7%, 15.8%, 9.2% and 4.4% for the category of spatial aspects. Dimensions higher than 2 in these two analyses do not carry interpretable information but plotting the correlation circles of PC #1 and PC #2 gives an informative picture of the individual attribute relationships. The relatively low explained variance covered with the two first principal components, that is 52.6% and 42.5% for the category of timbre and spatial aspects respectively, is partly due to the variability introduced by differences between the program items in the merged data sets considered in this study.

Figure 9.5(a) illustrates the correlation loading plot obtained for the category of timbral aspects. Two main attribute clusters are visible on this graph with *bass* attributes on the one side and other attributes related to timbral characteristics on the other side. These clusters might be described as two opposite perceptual directions associated with *low-frequency emphasis* and *high-frequency emphasis* respectively. The fact that the attribute *sharpness* of the assessor #9 correlates with the *bass* attributes in this graph can be attributed to an inversion in polarity as this term is semantically closer to the *high-frequency emphasis* cluster. Finally, it is interesting to note that all three occurrences of the attribute *muffling* are grouped to some extent and correlate moderately well with the second PC in this graph.

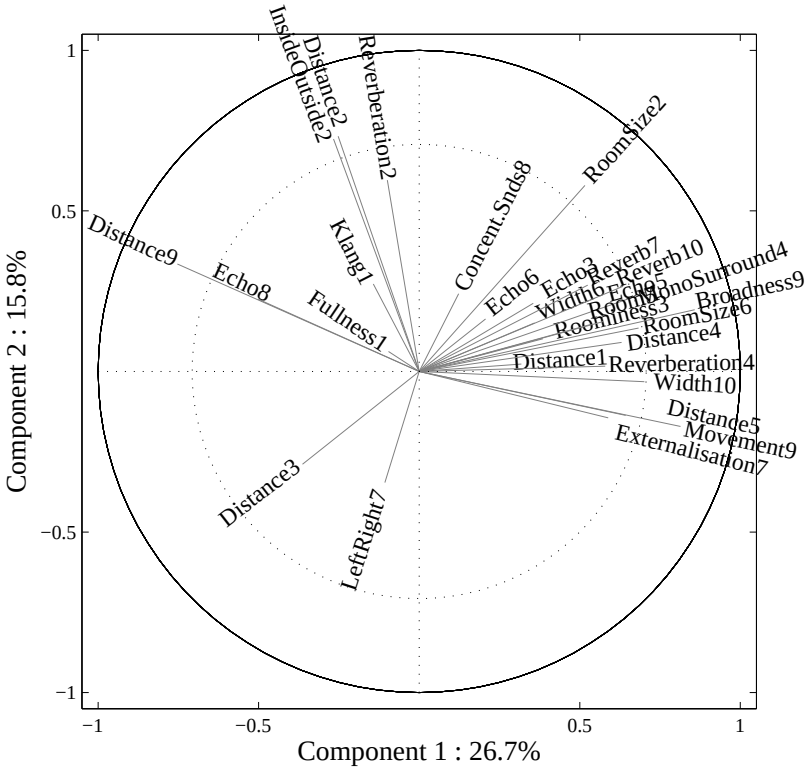
Figure 9.5(b) shows the correlation loading plot obtained for the category of spatial aspects. The main information in this graph can be seen along the first principal component with a single cluster of attributes relating to spatial characteristics. Few attributes correlating with the second dimension are visible but the identification of a perceptual cluster is difficult because of the unrelated descriptor names. It appears from this graph that the descriptor *distance* does not relate to a unique sensory concept as can be seen from the absence of correlation between few occurrences of this attribute.

Conclusion of the individual attribute grouping

The attribute classification attempted in this section illustrates the potential and limitations of the three techniques considered for this type of analysis. It became apparent by comparing the qualitative classification to the AHC and PCA results



(a) PCA on attributes relating to timbral aspects



(b) PCA on attributes relating to spatial aspects

Figure 9.5: PCA of the two main perceptual groups. These correlation loading plots illustrate the relationships between individual attributes and principal components. The inner and outer circles indicate 50% and 100% explained variance respectively.

that the level of details seen in Table 9.2 exceeds the true perceptual complexity of the IVP data set. This discrepancy is manifest when comparing the four qualitative groups relating to spatial aspects and the single perceptual dimension identified in the PCA of the full group of spatial attributes. Also, as for the comparison of the two quantitative methods for attribute grouping, PCA was found to be more suitable than AHC for the identification of sensory concepts having a similar nature but an opposite rating scale polarity.

9.4 Multivariate analysis per music clip

9.4.1 Application of the generalized Procrustes analysis procedure

After the initial exploration of the individual attributes and perceptual groups presented above, the focus in this section is placed on describing and quantifying the differences between the spatial enhancement algorithms under study. In Chapter 7, several multivariate analysis techniques were presented for the treatment of individual vocabulary data and the most popular technique referred to as 'generalized Procrustes analysis' was described thoroughly. GPA was therefore selected for the present analysis. Since the panel of ten assessors evaluated the eight systems under study for three music clips (Figure 9.2), a separate analysis was applied to the set of ten individual sensory configurations obtained for each music clip. In practice, the following three-step procedure was applied for the analysis.

- Application of GPA to the ten raw individual configurations using an implementation of the GPA algorithm written by the author in the Matlab software package (see Appendix C for details).
- Validity assessment of the GPA group average with the statistical test proposed by Wakeling *et al.* (1992). The idea of this procedure is to apply a random permutation to the rows of each individual configuration X_i to break the relationships between samples across assessors. When a GPA is run on this permuted set of data, the proportion of the total variance explained by the GPA-average (R_c) represents the chance level. By repeating this procedure a large number of times (e.g. 5000), a distribution of R_c is drawn from which the 95^{th} percentile can be observed (U^* following the notation of Wakeling *et al.*, 1992). The proportion of variance accounted for by the GPA-average obtained from the original sensory data set (the 'empirical' R_c) can then be compared to the value of U^* to check its significance level. An example of output obtained from this procedure is illustrated in Figure 9.6 for the sensory data set of the music clip #2. The distance between the two values in this case indicates that the GPA average is statistically significant.
- Application of PCA to the group average configuration to inspect the differences between systems and study the relationships between the original individual attributes in a space of lower dimensionality.

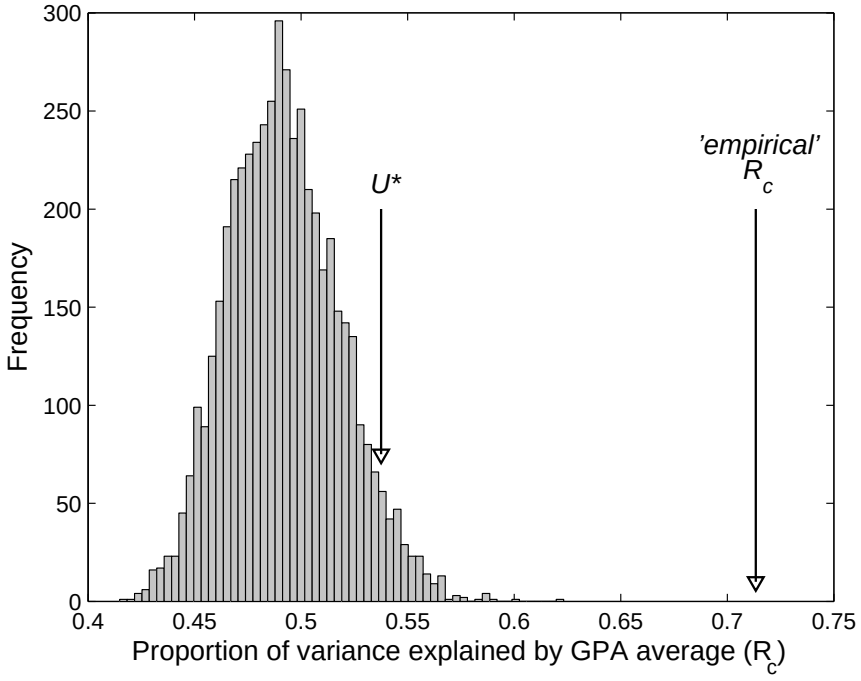


Figure 9.6: Illustration of the statistical procedure proposed by Wakeling *et al.* (1992) to test the level of significance of a GPA group average. In this example, the distribution of R_c is based on 5000 random permutations of the samples from the music clip #2 data set.

The visualization of the PCA results follows the principle of score plots and correlation loading plots already applied earlier in this thesis as illustrated for the music clip #1 in Figure 9.7 and Figure 9.8 respectively. Note that the graphs relating to the two other music clips are shown in Appendix B.

The score plots represent the panel average scores of the eight audio systems and include for each system an ellipse illustrating the 95% confidence level for the associated mean. This technique provides a way to visualize significant perceptual differences between systems at a panel level. The ellipses were computed from the individual GPA-transformed data using the bootstrapping method of Husson *et al.* (2005).

A correlation loading plot approach was adopted to illustrate the relationships between each individual attribute and the principal components. The readability of this type of plots in GPA applications is usually limited by the large number of individual attributes to be displayed but subsets of attributes can be shown in separate correlation circles to facilitate the analysis of attribute relationships. In the present study, the attributes relating to timbral and spatial aspects were separated following the classification presented earlier (see Table 9.3) while the 11 remaining attributes were included in all graphs. A visual interpretation was applied for each music clip by studying the level of correlation between individual attributes of a given perceptual group or sub-group and then adding arrows on the correlation loading plots to highlight the main underlying perceptual directions identified. Finally, this information was utilized to explain the sample differences observed in the associated score plots in terms of sensory characteristics.

9.4.2 Result for the music clip #1: *Scritti Politti* track

The group average configuration obtained from the GPA routine applied to the ten sensory configurations of the music clip #1 was tested with the permutation procedure proposed by Wakeling *et al.* (1992) and was found to be statistically significant ($R_c = 0.636$, $U^* = 0.601$, 5000 random permutations). The PCA model applied to the GPA average configuration explained 84.2% of the variation in the data with 4 components (PC #1 = 34.2%, PC #2 = 24.1%, PC #3 = 14.7%, PC #4 = 11.2%). Figures 9.7 and 9.8 present the score plot and correlation loading plots of the two first components. The information contained in the score plot of the third and fourth dimensions is difficult to interpret due to the low correlation loading values of the individual attributes. These principal components are therefore left out from the present analysis. The level of separation between the ellipses of the score plot (Figure 9.7) indicates a significant panel discrimination between the eight systems for this music clip. The algorithm map can be broadly summarized in three distinct areas defined by the HRTF-based algorithm 6_{Se} and the unprocessed mono 2_{M} in the left part of the score plot, the algorithm 8_{Se} and the mono-to-3d algorithm 4_{Me} in the bottom-right quadrant and the system 5_{Se} in the upper part of the score plot.

The correlation loading plots of Figure 9.8 can be exploited to identify the important perceptual dimensions of this set of sensory profiles. It appears from Figure 9.8(a) that six (out of nine) *bass* attributes are clustered and correlate well with the second principal component. The other timbral attributes present in this graph carry less consistent information except for a small group of five timbral attributes (including the inverted attribute *Sharpness* of the assessor #9) correlating positively with PC #1 and negatively with PC #2. Looking now at the Figure 9.8(b), a group of ten spatial attributes correlating positively with PC #1 can be highlighted. The perceptual directions identified from this two-dimensional space are summarized by the three arrows labeled *low-frequency emphasis*, *timbral aspects* and *spatial aspects*.

A combined analysis of the Figures 9.7 and 9.8 gives the following elements of interpretation for this music clip:

- the system 5_{Se} shows more low-frequency emphasis than the other systems,
- the systems 4_{Me} and 8_{Se} are characterized positively in terms of timbre,
- the systems 6_{Se} and 2_{M} are characterized negatively in terms of spatial aspects while the systems 1_{S}, 5_{Se} and 8_{Se} are characterized positively on this direction.

9.4.3 Result for the music clip #2: *Madonna* track

The GPA group average obtained for the ten sensory configurations of the music clip #2 was also found to be statistically significant ($R_c = 0.713$, $U^* = 0.537$, 5000 random permutations, see Figure 9.6) and the PCA model applied to the resulting configuration explained 89.7% of the variation in the data with 4 components (PC #1 = 50.0%, PC #2 = 17.9%, PC #3 = 13.7%, PC #4 = 8.7%). Figures D.1 and D.2 in appendix present the score plot and correlation loading plots of the two first components, which carry the most relevant information. The relatively low level of overlap between the ellipses in the score plot (Figure D.1) indicates that the panel

also perceived significant differences between the eight systems for this second music clip. However, a slightly different sensory map can be observed in this case with three following areas: the algorithm 5_{Se} and the unprocessed stereo 1_S in the upper part of the score plot, the algorithm 6_{Se} in the bottom-right quadrant and the algorithm 8_{Se} in the bottom-left quadrant of the plot.

Looking at the correlation loading plots shown in Figure D.2, a perceptual grouping similar to the previous music clip can be identified. Figure D.2(a) highlights an opposition between two clusters of timbral attributes along PC #1 with one group of eight *bass* attributes on the right side and a second group of ten timbral attributes on the left side. A large cluster of spatial attributes correlating negatively with PC #1 can also be seen in Figure D.2(b) but with a relatively wide spread along PC #2. Finally, it should be noted that the distribution of individual attributes along PC #2 is not systematic across the panel in these two graphs, which can be interpreted as an absence of consensus between assessors on this principal component.

By combining the information from the score plot in Figure D.1 and the arrows in the two Figure D.2, the following elements of interpretation can be listed for this music clip:

- the systems 6_{Se}, 3_{Se}, 2_M and 5_{Se} show more low-frequency emphasis than the other systems,
- the system 8_{Se} is characterized positively in terms of timbral aspects,
- the system 6_{Se} is characterized negatively in terms of spatial aspects while the system 8_{Se} is characterized positively on this perceptual direction.

9.4.4 Result for the music clip #3: *Tuck & Patti* track

To complete this separate study of the three music clips, the data set obtained for the music clip #3 was analyzed in a similar way. The group average configuration obtained from the GPA was found to be statistically significant again ($R_c = 0.572$, $U^* = 0.479$, 5000 random permutations) but it can be noted that the variance explained by this mean configuration is the lowest of the three music clips. The PCA model in this case explained 87.1% of the variation in the data with 4 components (PC1 = 47.0%, PC2 = 19.3%, PC3 = 12.5%, PC4 = 8.3%) and only the two first components were considered for the PCA interpretation as illustrated in Figures D.3 and D.4 in appendix. The ellipses of the score plot (Figure D.3) show a sensibly larger size than in the two previous analyses, which can be interpreted as a lower agreement at a panel level. A reduced discrimination level results as illustrated by the overlap between the ellipses of the unprocessed stereo 1_S and the two algorithms 5_{Se} and 6_{Se}.

Considering the correlation structure obtained from this PCA, attribute clusters similar to the previous music clips can be highlighted. From Figure D.4(a), it appears that a group of eight timbral attributes correlate well with PC #1 and a group of the nine *bass* attributes correlate to a lower extent with PC #2. Also, eleven spatial attributes showing a large correlation loading value form a tight cluster in the bottom-right quadrant of Figure D.4(b).

Based on this correlation circle analysis, the sensory differences between systems seen in Figure D.3 can be interpreted as follows:

- the system 3_{Se} shows more low-frequency emphasis than the other systems,
- the systems 8_{Se} and 4_{Me} are characterized positively in terms of timbral aspects,
- the system 2_{M} is characterized negatively in terms of spatial aspects.

9.4.5 Synthesis of the generalized Procrustes analysis results

The results of the GPA procedure applied to each music clip can be compared and eventually combined at a qualitative level to highlight the global features of the spatial enhancement algorithms evaluated in this study.

It should be noted first that a valid panel sensory map was obtained for the three music clips as illustrated by the significant GPA-group average configuration obtained in each case and by the significant level of panel discrimination between audio samples resulting from the PCA subsequently applied to the GPA-transformed data.

This exploratory analysis also showed that the 2-dimensional PCA maps interpreted above share some similarities across music clips both in terms of differences between systems and associated perceptual description. More specifically, the two main perceptual groups considered in this GPA study appear to be represented (to some extent) in a 'consensual' way in the three analyses and two separate perceptual directions relating to timbral aspects were also identified, that is, *low-frequency emphasis* including *bass* attributes and a less specific group including the other timbral attributes. However, it can be noted that the correlation structure observed between these three perceptual clusters is less systematic across music clips since a high negative correlation between the two timbral sub-groups is only visible for the music clip #2 and a somewhat expected de-correlation between the spatial and timbral dimensions is manifest for the music clip #1 but is not clear at all for the music clip #2.

Concerning the attributes not part of the two main perceptual groups which were included in all the correlation loading plots presented above, it should be noted that some of them fit the identified clusters at several occasions. For example, the attribute *Sound Impression* of Assessor #3 was found to correlate with the group of *Bass* attributes for the music clips #1 and #2. However, the other attributes do not show a systematic pattern and/or they can not be interpreted reliably because they are under-represented.

To summarize the results of the three sets of analyses presented above, a global interpretation of the main sensory differences between the spatial enhancement algorithms under study can be given as follows:

- **the systems $3_{\mathbf{Se}}$ and $5_{\mathbf{Se}}$ show more low-frequency emphasis than the other systems**,
- **the systems $8_{\mathbf{Se}}$ and $4_{\mathbf{Me}}$ are characterized positively in terms of timbral aspects**,
- **the systems $2_{\mathbf{M}}$ and $6_{\mathbf{Se}}$ are characterized negatively in terms of spatial aspects while the system $8_{\mathbf{Se}}$ is characterized positively on this perceptual direction**.

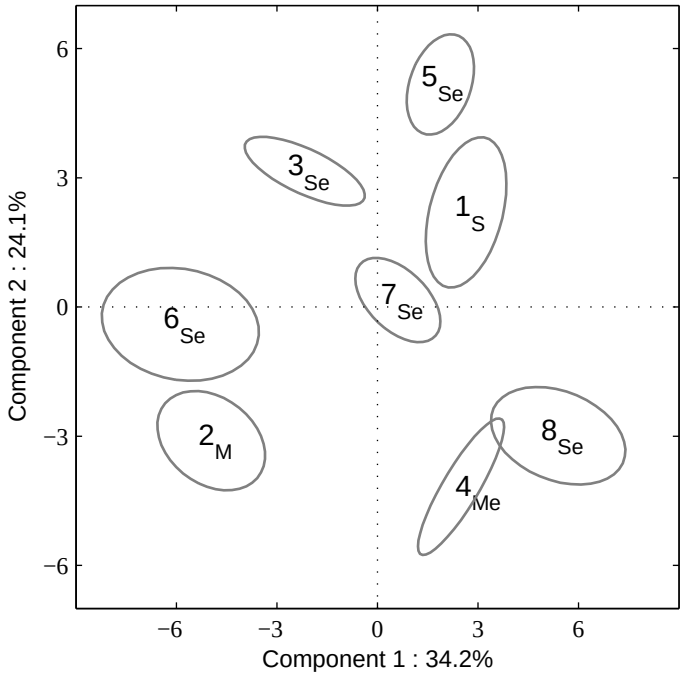
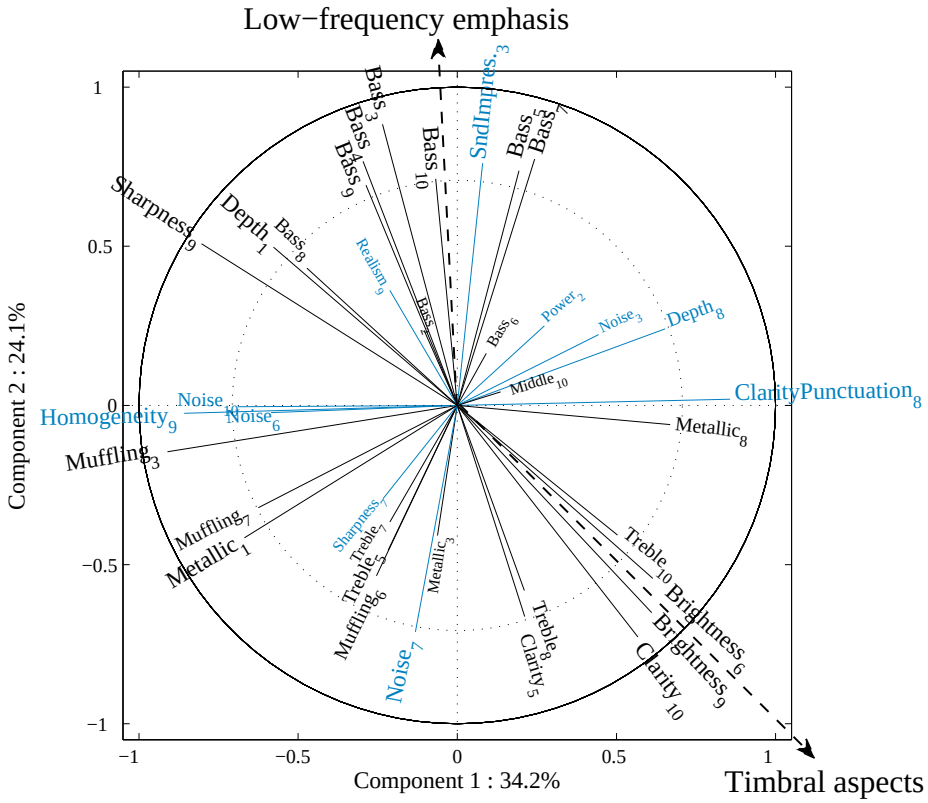
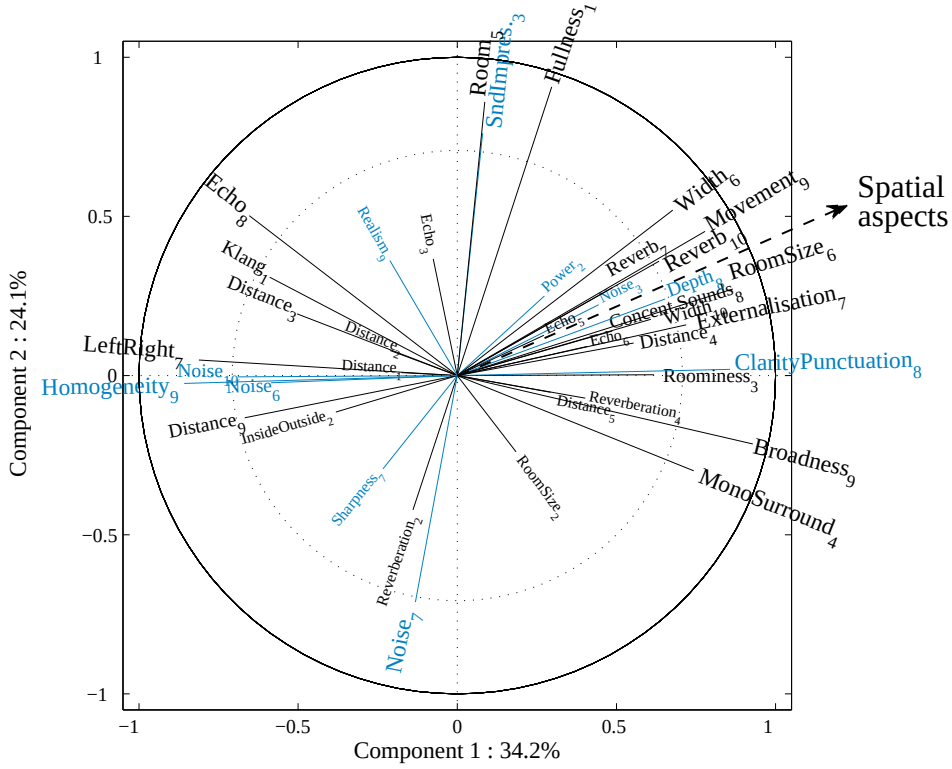


Figure 9.7: PCA analysis of the GPA average configuration obtained for the music clip #1 – Score plot.



(a) Timbral aspects



(b) Spatial aspects

Figure 9.8: PCA analysis of the GPA average configuration obtained for the music clip #1 – Correlation loading plots.

9.5 Multi-way analysis of the full data set

The analysis presented above showed that the perceptual differences between the spatial enhancement systems under study can be explored in detail by an application of the GPA-PCA procedure to each music clip separately. However, one limitation of this approach is that the interpretation of sensory characteristics across music clips was only possible at a qualitative level. One way to overcome this issue is to consider a combined analysis of the three sets of sensory profiles in order to be able to measure quantitatively the global differences between systems, to identify the associated perceptual dimensions and to estimate the possible sensory differences between music clips. This could be achieved by several multivariate data analysis techniques but the focus of the present section is placed a novel application of parallel factor analysis 2 (PARAFAC2) to the four-way individual vocabulary data set under study.

9.5.1 Representation of the full data set as a 4-way array

The full sensory data set of the spatial enhancement algorithm study illustrated in Figure 9.2 comprises 30 sensory profiles and can be represented with the four modes `System`, `Attribute`, `Assessor` and `MusicClip`. Note that the modes `System`, `Assessor` and `MusicClip` are 'fixed' in the sense that the same eight systems were evaluated by all ten assessors for all three music clips. However, assessors employed their own set of sensory descriptors to assess the systems for the three music clips, which implies that the mode `Attribute` varies across the mode `Assessor`. A visual representation of this sensory data set as a four-way array is shown in Figure 9.9.

It should be noted that the use of a comparative approach in this sensory experiment and the fact that assessors were instructed to focus on relative differences between systems reduce the ability of the collected data to capture sensory differences between music clips in an absolute sense. Nevertheless, the proposed combined analysis still allows to compare relative differences between algorithms across music clips and therefore find out whether a specific system exhibits more of a certain sensory characteristic than the other systems for all three music clips or for only some of them.

Several approaches are available to analyze this type of data structure. As the principal interest in the present application is to study the sensory characteristics of the systems at a global level the following 'unfolding' method might be considered. First, the mode `MusicClip` is unfolded on the mode `Assessor` to obtain a set of 30 (10×3) sensory profiles and then any of the multivariate data analyses presented in Section 7.3 is applied to the resulting data set. The drawback of this approach is that sensory information specific to a certain music clip is not directly taken into account. Alternatively, two multivariate data analysis techniques that can handle explicitly the mode `MusicClip` have been illustrated by Lorho (2008), namely, Hierarchical Multiple Factor Analysis (HMFA, Le Dien and Pagès, 2003a,b) and PARAFAC2 (Harshman, 1972). The latter approach has been selected for the present study because it is well suited but not often applied to sensory data. For more details about this multi-way analysis technique, the reader is directed to the introduction of PARAFAC2 provided in Section 7.3 and to the mathematical presentation given in Appendix E.

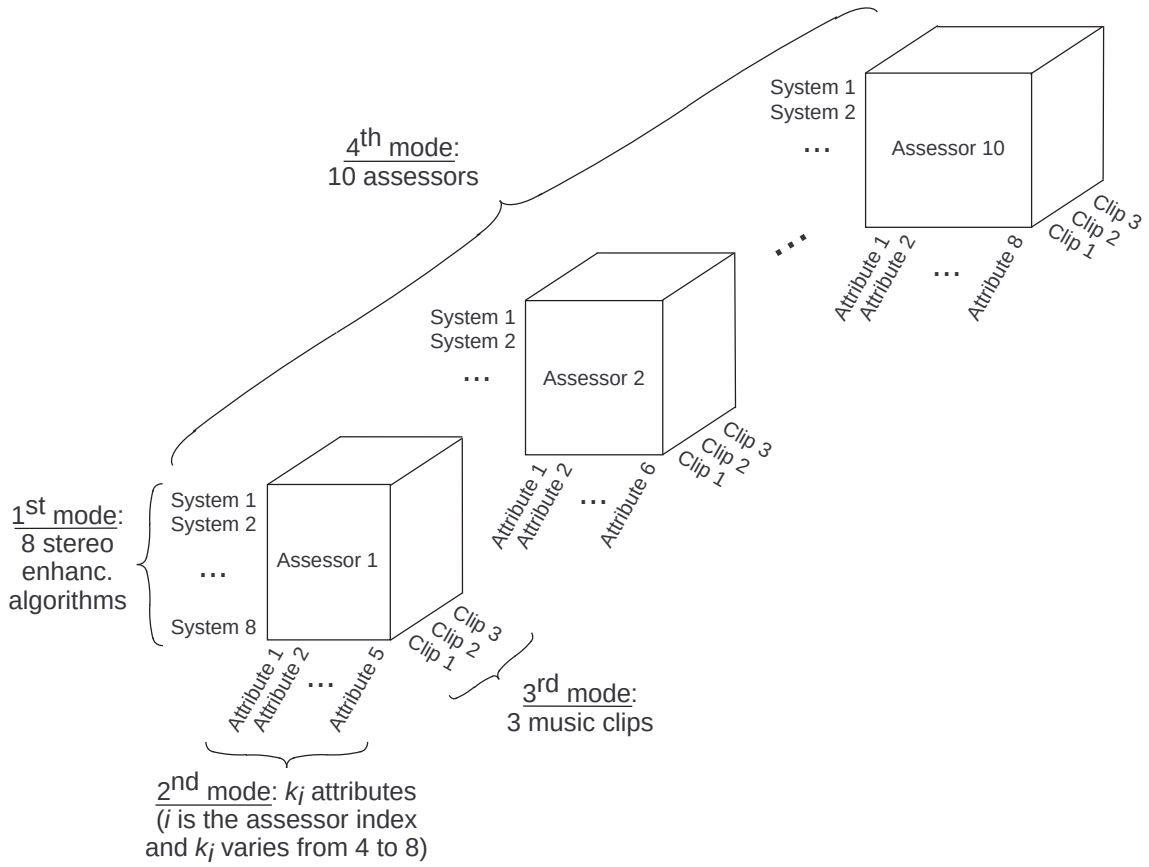


Figure 9.9: Representation of the sensory data set considered in this study as a four-way array fitting the PARAFAC2 modeling approach.

9.5.2 Data modeling by parallel factor analysis 2

The analysis reported in this study was performed with the PARAFAC2 algorithm of the PLS toolbox 4.0 (2006) running in the Matlab software environment. The raw four-way data set was first pre-processed by centering the array across the mode `System` and scaling it within the mode `Assessor`. As already described in Chapter 6, the procedure of scaling 'within a mode' does not affect the structure of the data at a given level in the mode considered, which means in the present case that scaling differences between attributes and between music clips are preserved for each assessor. The aim of this scaling procedure is only to balance out the contribution of the assessors in the analysis.

A two-component solution of the pre-processed data was selected as it seemed the most appropriate although it explains only 38.47% of the variation in the data. It should be noted that this relatively small value can not be compared directly to the explained variance obtained for the GPA presented in the previous section because PARAFAC2 models directly the (pre-processed) data while GPA explained variance figures relate to the PCA model of the GPA average configuration. As a matter of fact, combining the unexplained variance of the GPA and PCA steps yields an averaged

explained variance of 40.1% in the previous section, which is comparable to the present PARAFAC2 model results.

The output of this two-component PARAFAC2 model comprises a set of loadings for the fixed modes `System`, `Assessor` and `MusicClip` shown in Figure 9.11 and a set of loadings per assessor for the 'free' mode `Attribute`. The description of the model output provided below starts with the mode `Attribute` because the PARAFAC2 latent components offer a relatively simple sensory interpretation and the subsequent description of the loadings of the three other modes can then be performed in an intuitive way based on the interpreted perceptual dimensions.

Also, while the three fixed modes are interpreted directly from their loadings, the large number of model parameters resulting from the free mode `Attribute` is visualized with a multi-way equivalent of the correlation loading plot found in PCA applications (see Figure 9.11) which is referred to as the 'congruence loading' plot (Lorho *et al.*, 2006). Two specificities need to be taken into account when applying this type of plot to PARAFAC models. Firstly, variables from which correlation loadings are derived do not always relate to a centered mode and for this reason congruence loadings, that is, uncentered correlation loadings, are computed. Secondly, loadings of different components are not orthogonal in PARAFAC models, so that the variation explained by two components is not additive and each component should be assessed individually. The use of circles in congruence loading plots is therefore not meaningful as they do not imply 100% explained variance and for this reason plots are shown with squares rather than circles in Figure 9.10. Note also that to facilitate the interpretation of attribute clusters, a split between the two main groups of timbral attributes and spatial attributes was applied for the plotting of these congruence loadings as already applied the in GPA study presented earlier.

9.5.3 PARAFAC2 model interpretation

Individual attributes

Focusing first on the Figure 9.10(a), it can be seen that the component #1 of the PARAFAC2 model relates almost exclusively to the group of timbral aspects as highlighted by the two ellipses added to this graph. Seven out of nine *Bass* attributes and the attribute *Depth* of Assessor #1 are clustered on the left side of the congruence loading square while a group of timbral attributes is also clustered on the other side. Based on the opposition between these two clusters, this component can be interpreted as a *frequency emphasis* dimension with the two opposite polarities of *high-frequency emphasis* in the positive direction and *low-frequency emphasis* in the negative direction. In light of this clustering, the *Sharpness* attribute of Assessor #9 appears to have an inverted polarity, as already noted from the GPA study, and the *Fullness* attribute of Assessor #1 seen in Figure 9.10(b) seems to fit well the group of *low-frequency emphasis* attributes, which indicates a potential miss-classification of this individual attribute.

Looking then at the Figure 9.10(b), the second component of the PARAFAC2 model appears to relate largely to the group of spatial aspects. At least twelve in-

dividual attributes of this group are clustered in the lower part of the congruence loading 'square' as highlighted by the ellipse added to the graph and all these terms refer to a positive concept of spatial sound perception, i.e., *width*, *surround*, *room effect*, *reverberation*, *broadness*, *externalization*, etc. This component can therefore be interpreted as a *spatial emphasis* dimension indicating a small or a large amount of spatial characteristics in the positive or the negative direction respectively.

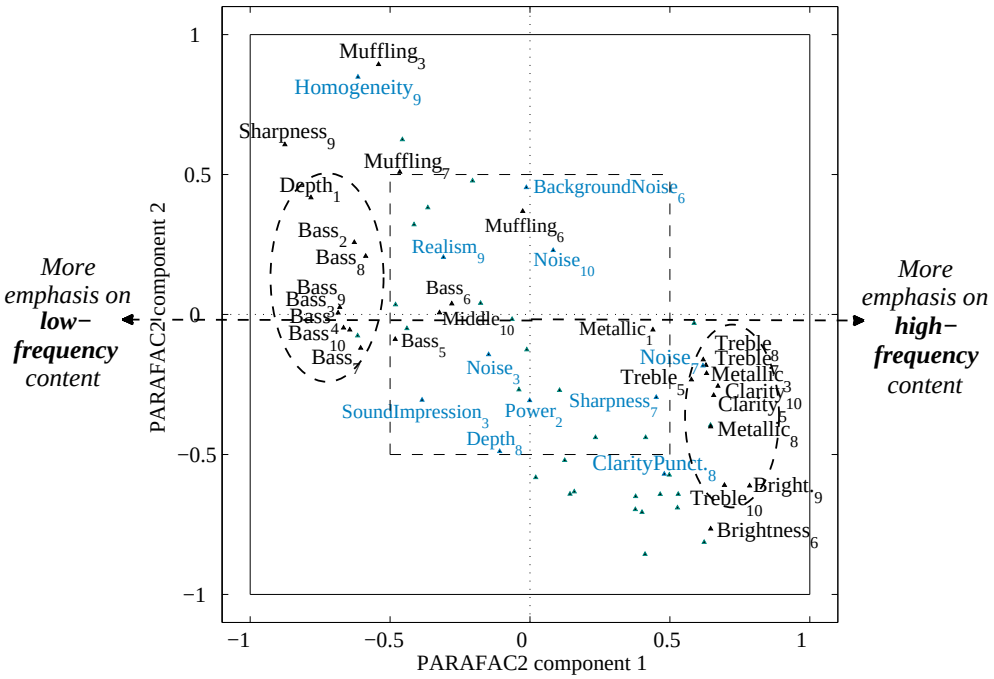
Three comments can be made to complete the review of the Figure 9.10. Firstly, the congruence loading of few timbral and spatial attributes appear to be high for both components, e.g. the attribute *Broadness* of Assessor #9 and the attribute *Brightness* of Assessor #6. By studying the congruence loadings of the mode `System` (not shown here), it appears that only the system 8_{Se} has a relatively large value (≥ 0.6) on both components which indicates that this timbral/spatial attribute co-variation is mainly due to that spatial enhancement algorithm. Secondly, two occurrences of the attribute *Muffling* can be seen in the upper part of the congruence loading square despite the fact that they are not directly related to a spatial characteristics. Thirdly, the five occurrences of the attribute *Distance* appear at very different positions in the Figure 9.10(b), which indicates that they do not relate to the same sensory concept.

Spatial enhancement algorithms

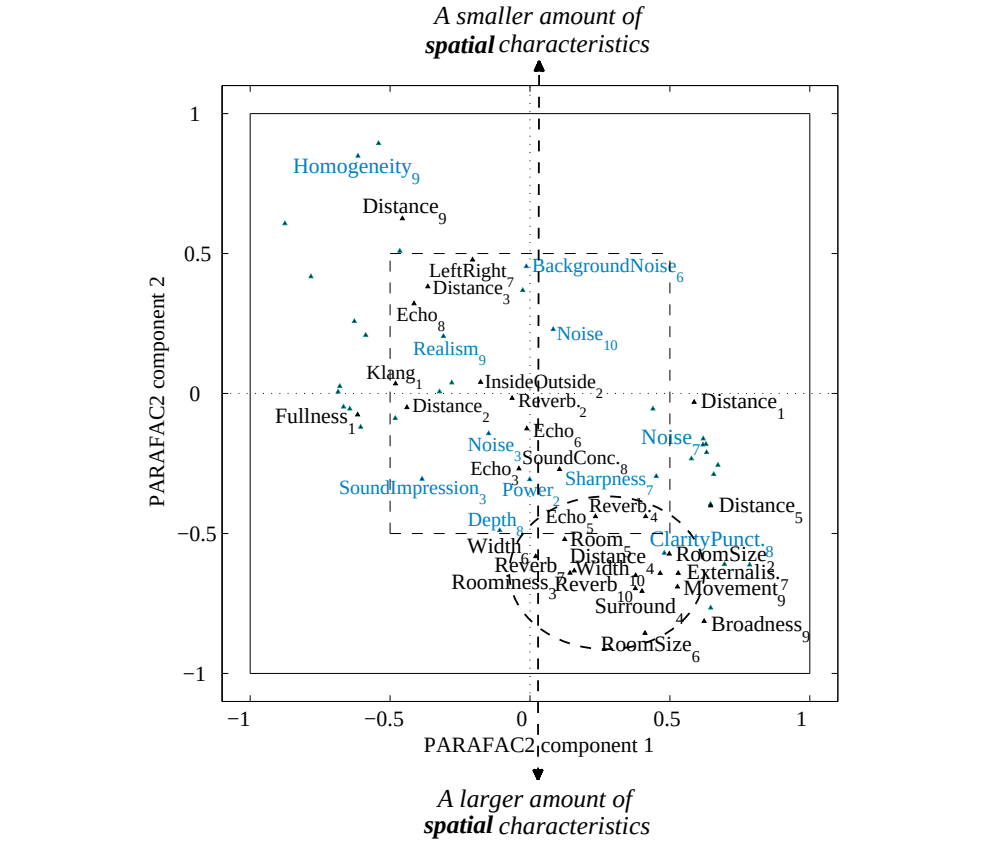
The upper pane of Figure 9.11 illustrates the loadings of the mode `System`. Based on the component interpretation presented above, the position of the spatial enhancement algorithms in this graph can be analyzed directly in terms of *frequency emphasis* on the x-axis and *spatial emphasis* on the the y-axis. Looking at the component #1, the stereo (1_{S}) and mono (2_{M}) reproduction techniques can be seen to be neutral in terms of *frequency emphasis*, which is an expected result overall. In addition, two groups of systems are visible along this dimension with the three systems 3_{Se}, 6_{Se} and 5_{Se} emphasized on the low-frequency side and the two systems 4_{Me} and 8_{Se} emphasized on the high-frequency side emphasis. On the component #2, a clear separation can be seen between the two systems 2_{M} and 6_{Se} having a small amount of spatial characteristics and the other systems. It also appears that the system 3_{Se} is neutral spatially while the spatial enhancement algorithm 8_{Se} has the largest amount of spatial characteristics.

Assessors

The loadings of the `Assessor` model shown in the middle pane of Figure 9.11 represent the weight given by each assessor for the two components, i.e., the extent to which they use the timbral dimension on component #1 and the spatial dimension on component #2 for discriminating the systems. It can be noted first that all these loadings are positive, which indicates that no major rating inconsistency exists between assessors. It also appears from this graph that half of the assessors employ equally the timbral and spatial dimensions as can be seen from the grouping around 0.3 on each axis, while the other assessors have a somewhat different sensitivity towards the two perceptual dimensions. The assessors #1 and #9 can be considered extreme in the sense that they are over-represented on one of the component while the assessor #1 seems to be



(a) Timbral aspects



(b) Spatial aspects

Figure 9.10: PARAFAC2 model - Congruence loading plot of the mode `Attribute` split in two attribute groups relating to timbral aspects (a) and spatial aspects (b).

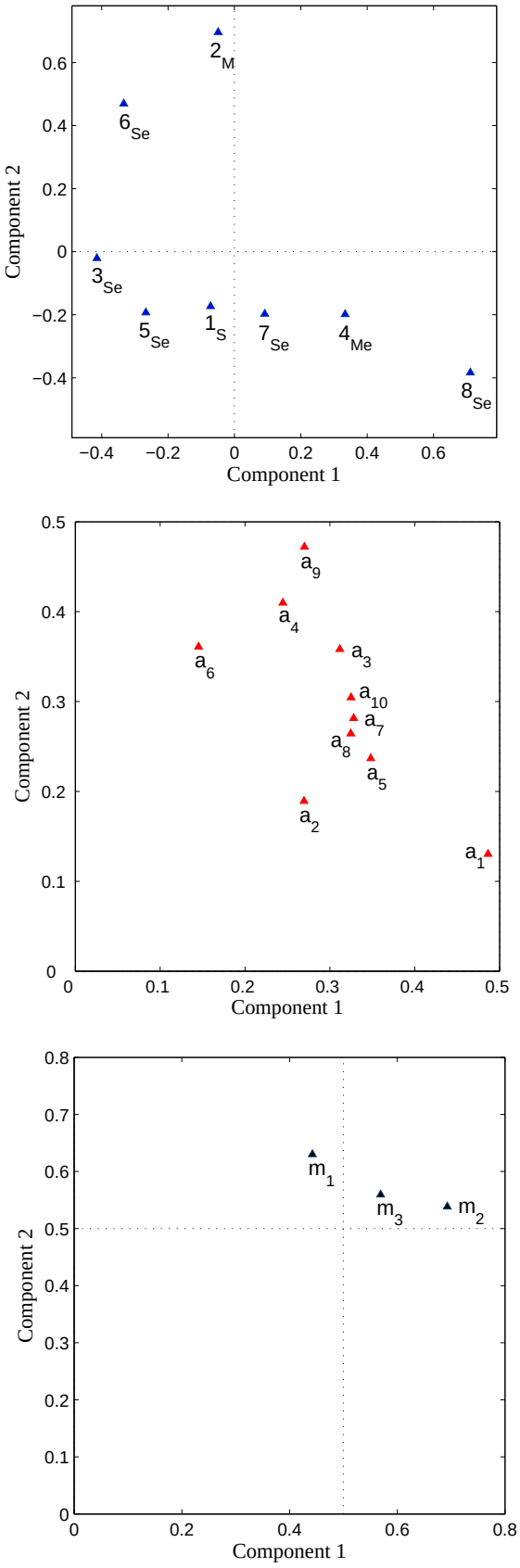


Figure 9.11: PARAFAC2 model - Loading plot of the modes `System` (upper pane), `Assessor` (middle pane) and `MusicClip` (lower pane).

unable to perceive the timbral aspects, at least in a consensual way.

Music clips

The loadings of the `MusicClip` mode shown in the lower pane of Figure 9.11 describe the extent to which each music clip is represented on the two latent components. This graph illustrates a slight difference between the three clips as the music clip #2 is better represented on the timbral dimension while the music clip #2 is a bit better represented on the spatial dimension.

9.5.4 Summary of the PARAFAC2 analysis

An illustration of the PARAFAC2 model application to a four-way individual vocabulary data set was provided in this section. The sensory data resulting from the IVP of the spatial enhancement systems appeared to be well suited to this type of multi-way data analysis. The formal representation and modeling of this data set as a four-way structure provided a separate set of loadings for each mode, which were easily interpreted. While handling effectively the individual nature of the sensory descriptors, PARAFAC2 also offered a global and simple description of each assessor's vocabulary in terms of latent dimensions. It should be noted that unfolding analysis approaches mentioned earlier would produce a much larger number of model parameters, that is, one loading per attribute and per music clip in the present case, which can be a limiting factor in the model interpretation.

The semantic interpretation of the latent dimensions in this model highlighted a relatively clear separation between the two main perceptual aspects, that is, the timbral characteristics on the first component and the spatial characteristics on the second component. This straightforward mapping allowed for a simple sensory characterization of the eight spatial enhancement systems at a global level, which complemented the outcome of the GPA-PCA study presented earlier. Additionally, this analysis provided relevant information about the ability of each assessor to perceive the two perceptual dimensions and highlighted characteristic differences between music clips.

To complete this section on the application of multi-way models to sensory data, the following comment on the aspect of PARAFAC2 model validation can be made. While the outcome of the present model illustrated well the potential of this analysis technique, the author experienced problems of model interpretation with other sensory data sets and noted that model validation is a crucial step in PARAFAC2 that would deserve further investigation.

9.6 Conclusion

An application of the IVP procedure was illustrated in this chapter. The rapid descriptive evaluation of this set of spatial enhancement systems demonstrated that a rich sensory characterization is possible through verbal elicitation. The review of the large amount of data generated by this type of experiment highlighted the importance of a systematic analysis of results. This chapter gave an illustration of the different analysis tools available for this purpose. First, the qualitative and quantitative

exploration of individual attributes was applied to identify and validate perceptual groups. Then a quantitative analysis of the individual profiles was carried out with two different multivariate data analysis techniques. The application of the GPA-PCA procedure to each music clip produced reliable results and the PARAFAC2 multi-way analysis method was proposed to take into account the four-way structure of this type of sensory data set typical to the field of audio quality evaluation.

Chapter 10

Comparison of the consensus vocabulary and individual vocabulary approaches

10.1 Introduction

In the previous chapters, a comprehensive description of the two different verbal elicitation approaches referred to as 'consensus vocabulary' and 'individual vocabulary' was provided (Chapters 4 and 7) and their application to the same set of spatial enhancement systems for headphone reproduction was illustrated (Chapters 6 and 9). To summarize this twofold investigation, a comparative analysis of the two approaches is proposed in two steps. Firstly, a combined analysis of the CV and IV data sets is applied to assess the level of similarity between the results of these two experiments. Secondly, a more general comparison of the two methodologies is provided highlighting their respective advantages, limitations and suitability for different sensory analysis projects.

10.2 Comparison of the CV and IV profiles of the spatial enhancement systems

In this section, the CV and IV approaches are compared from an application viewpoint by looking at the sensory profiling data resulting from the two experiments on spatial enhancement systems reported in Chapters 6 and 9. After providing some justifications to the application of this type of comparison, the principle and application of the hierarchical multiple factor analysis (HMFA) technique applied for this comparative study are presented and the results of this quantitative data analysis are discussed.

10.2.1 Scope of the comparative analysis

The procedure applied to obtain a quantitative sensory description of the set of spatial enhancement systems under study differed considerably between the two studies

reported earlier. While the CV experiment employed a generic vocabulary developed from a large set of headphone sound stimuli, the IV experiment employed exclusively the set of spatial enhancement systems for the elicitation process. It should also be noted that a slightly different type of sensory panel was selected for the two experiments. The CV experiment made in Finland employed selected assessors who developed the set of consensus attributes while the IV experiment made in Denmark employed six initiated assessors and four naïve assessors.

Despite these differences, the sensory profiling data resulting from the two experiments has a very similar format and comprises a large set of attribute rating scores. On the one hand, the CV data analyzed in Chapter 6 can be represented as three consensus sensory profiles, that is, a panel average configuration for each music clip (see lower row of Figure 6.3). On the other hand, the IV data analyzed in Chapter 7 can be represented as three sets of ten individual configurations (see Figure 9.2).

The idea of the present investigation is to assess the level of similarity between the sensory characterizations obtained in the two experiments, which both relate to the same entities, that is, the eight spatial enhancement systems under study. A combined analysis of these two sensory characterizations can be applied for comparison purpose and the present assumption is that any discrepancy found between the two data sets relates to the experimental procedure in a global sense.

Several meta-analyses of this type have been reported in the sensory science literature. Studies comparing sensory panels of food products have been performed for example on chocolate (Risvik *et al.*, 1992; Pagès and Husson, 2001), coffee (de Jong *et al.*, 1998; Schlich, 1998) and red wine (McEwan *et al.*, 2002). The investigation presented in this section followed the principle of these studies and employed a quantitative analysis technique inspired from the study of Le Dien and Pagès (2003b) comparing the sensory profiles produced by a 'trained' panel and an 'untrained' panel. This approach referred to as 'hierarchical multiple factor analysis' (HMFA) is presented next.

10.2.2 Hierarchical multiple factor analysis

Hierarchical multiple factor analysis (HMFA) is a flexible multivariate data analysis method introduced by Le Dien and Pagès (2003a) for the joint analysis of several sets of variables. This analysis technique extends the MFA approach introduced in Chapter 6 by taking into account the hierarchical structure of the data. In the example of the IV experiment (see Figure 9.2), a hierarchical structure would be created by highlighting either the music clips or the assessors. In both cases, the first level would consider each individual sensory profile separately as a matrix X_i with n rows of objects (the eight systems under study) and p_i columns of of centered variables (the attributes of the assessor i). However, the second level of the hierarchy would group the sensory profiles either by music clip in the former case or by assessor in the latter case.

Following the principle of MFA, the contribution of the different groups of variables is balanced at each node of the hierarchy before applying a PCA on the full matrix. HMFA offers graphical tools similar to PCA for the visualization of the objects and variables at the different levels of the hierarchy in terms of latent dimensions.

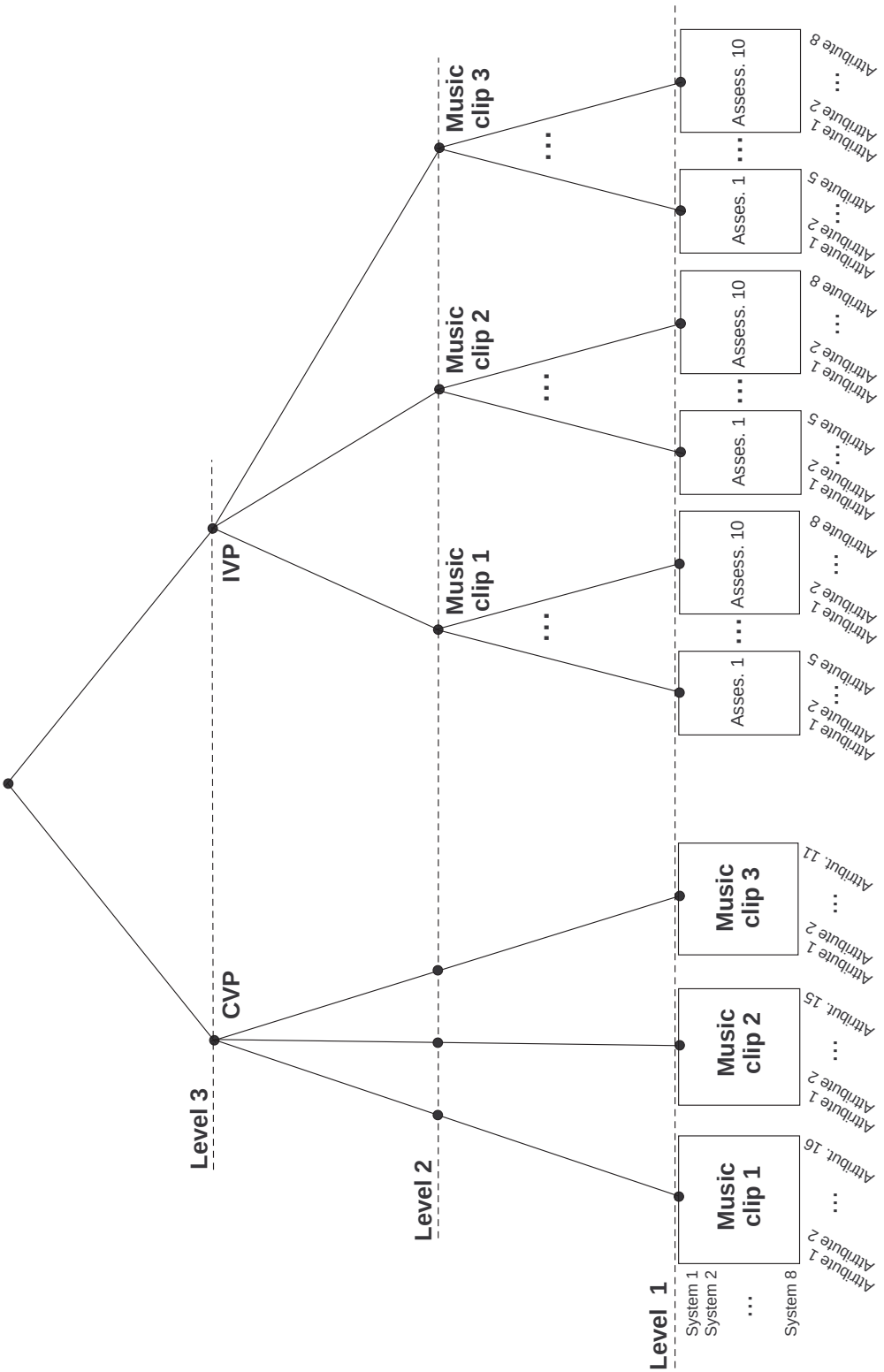


Figure 10.1: Hierarchical representation of the two experimental data sets.

The present analysis was performed with the HMFA algorithm of the FactoMineR package (Lê *et al.*, 2008) running in the R software environment. The option 'centering' was selected, which put to zero the mean of each variable separately.

The hierarchical structure employed for this analysis is illustrated in Figure 10.1. The highest level (Level #3) of the hierarchy separates the two experiments while the level below (Level #2) groups the data per music clip and the lowest level (Level #1) contains 33 separate sensory profiles. The CV data set includes three average sensory profiles made of 16, 15 and 11 attributes respectively[1] while the IV data set includes 30 (raw) individual sensory profiles adding up to a total of 197 attributes. The structure of the IV data set can be represented in a straightforward manner through this hierarchy but the CV data structure can be seen to have a redundant second level because the panel average data has been chosen for the present analysis. It can be noted however that another hierarchy could be considered in which the individual sensory profiles of the CV experiment are exploited instead of the panel average to obtain a hierarchical structure similar to the IV data set. This alternative approach implies that the data provided by the consensus panel is represented as a set of individual vocabulary profiles, which relaxes the assumption of a consensus between assessors. Both approaches have been investigated in this study but the only former HMFA structure is treated in detail below while the latter approach is discussed briefly at the end of this section.

10.2.3 HMFA results

The HMFA model applied to the hierarchical structure presented in Figure 10.1 explained 83.1% of the variation in the data with 4 components. The contribution of each component was 37.2%, 19.9%, 14.1% and 11.9% respectively while each of the remaining dimensions were found to have a smaller contribution (less than 6% explained variance) and were therefore not included in this analysis. The type of visualization tools found in MFA was employed for the interpretation of this HMFA model, that is, the global score plot, the correlation loading plot and the L_g plot representing the link between the group of variables and the different principal components (Escofier and Pagès, 1988). Additionally, the nodes defining the group of variables at different levels of the hierarchy were studied. In particular, the global scores representing the sensory map of the eight spatial enhancement algorithms across the two experiments and the partial scores of the level #2 representing the sensory map of the two experiments were displayed on the same graph as illustrated in Figure 10.2.

Interpretation of the global score plot

The coding scheme applied in the two experiments is recalled at this point. The systems are coded as 1_{S} for the stereo unprocessed material, 2_{M} for the monophonic down-mixed version of the stereo material, 4_{Me} for the spatial enhancement algorithm of monophonic music material and 3_{Se} to 8_{Se} for the five stereo enhancement algorithms. The music clips are coded as 'music clip #1' for the *Scritti Politti* track,

[1]These sensory profiles result from the pre-processing step applied for the multivariate analysis presented in Section 6.3.3 (Chapter 6).

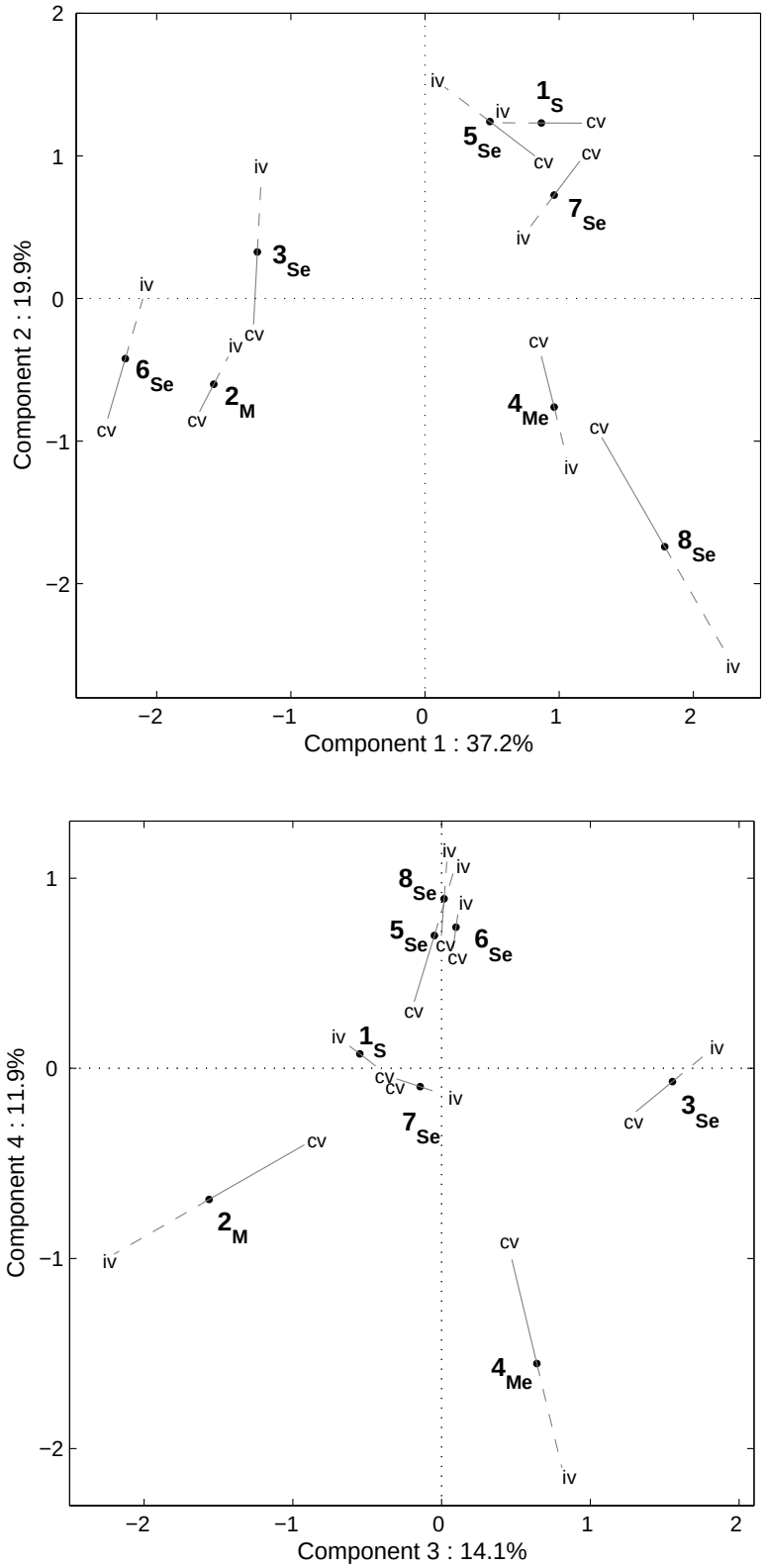


Figure 10.2: Score plot of the HMFA dimensions 1–2 (upper pane) and 3–4 (lower pane). The global scores represent the sensory map of the eight spatial enhancement algorithms across the two experiments (black dots) and the partial scores of the level #2 represent the map of each experiment (solid lines for CV and dashed lines for IV).

'music clip #2' for the *Madonna* track and 'music clip #3' for the *Tuck & Patti* track.

The global score plot is illustrated by the black dots in the two graphs of Figure 10.2 for the two first dimensions in the upper pane and for the third and forth dimensions in the lower pane. The grouping of systems seen in the upper graph appears to be similar to the result of the separate analyses reported in Chapters 6 and 9 for the two first dimensions. Three groups of systems can be identified, i.e., 1_S, 5_{Se} and 7_{Se} in the upper-right quadrant, 2_M, 3_{Se} and 6_{Se} on the left side and 4_{Se} and 8_{Se} on the lower-right quadrant. Additionally, the third and fourth dimensions of the HMFA model separate clearly the systems 2_M, 3_{Se} and 4_{Se} from the others systems.

Partial sensory map of the two experiments

The partial score plots of the second level are represented as deviations from the global score plots on the same graphs of Figure 10.2 with solid lines for the CV experiment and dashed lines for the IV experiment. These two scatter plots illustrate globally that the differences between the two experiments are smaller than the differences between systems, although relatively large variations can be seen for some algorithms, e.g. the system 8_{Se} in the upper graph and the systems 2_M and 4_{Se} in the lower graph. Interestingly, the group of three systems 1_S, 5_{Se} and 7_{Se} in the upper graph appears to be more clustered in the CV map than for in the IV map, which can be interpreted as a lower discrimination between the spatial enhancement systems for the CV experiment. It also appears from the lower graph that the CV map is globally more clustered than the IV map on the third and fourth HMFA dimensions.

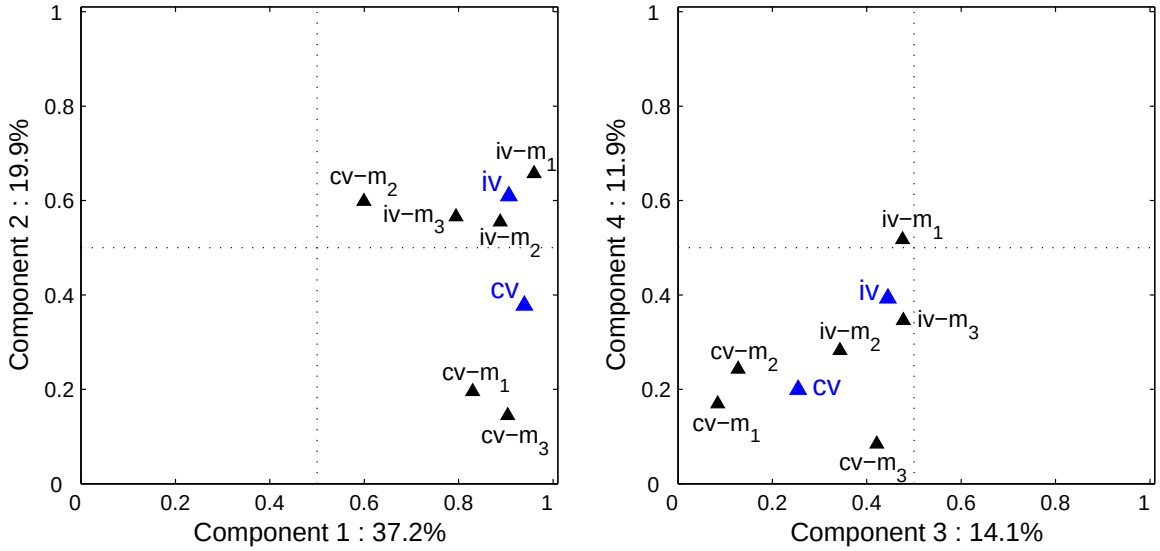


Figure 10.3: Representation of the groups of variables at the different nodes of the second and third level of the hierarchical structure. These L_g plots illustrate the contribution to the four HMFA dimensions of the two experiments (cv and iv) and the three music clips of each experiment (cv_m$_i$ and iv_m$_i$ with $i = 1, ..., 3$).

L_g plot

Figure 10.3 illustrates the link between the group of variables and the latent dimensions of the HMFA model. The large blue triangles correspond to the third level of the hierarchy separating the data sets of the CV and IV experiments while the small black triangles correspond to the second level separating the music clips within each experiment. Globally, the L_g index of all the groups of variables is higher on the two first dimensions (left plot) than on the dimensions three and four (right plot) and it appears that the IV experiment is better represented than the CV experiment on the second, third and fourth dimension. Additionally, a larger difference between the music clips can be seen for the CV experiment than for the IV experiment, especially on the three first dimensions. It should be noted that the difference observed between the music clip #2 and the two other music clips for the CV experiment are similar to the result of the MFA analysis presented in Chapter 6.

Semantic interpretation of the HMFA sensory map

As the principle of HMFA is to apply a PCA on a large matrix composed of weighted groups of variables, the number of parameters obtained in this type of model is usually large. The hierarchical structure employed in the present analysis contains 239 variables divided in 42 variables for the CV experiment and 197 variables for the CV experiment as illustrated in Figure 10.1. The correlation loading plot approach was adopted to visualize this large amount of data and identify possible clustering of attributes within or across experiments and music clips. In practice, a correlation circle was created for each group of variables at the level 2 and interpreted visually and perceptual directions were highlighted from this set of plots not presented here.

As a summary, it was found that the two first dimensions of the HMFA model relate to two main perceptual directions already identified in the separate experiments. A spatial direction was observed along the first dimension while a timbral emphasis direction emerged diagonally with two groups of attributes, that is, the low-frequency emphasis in the upper-left quadrant including *Bass* attributes mainly and the high-frequency emphasis in the lower-right quadrant including *Treble* attributes of the IV experiment and the attribute *Tone color* of the CV experiment). The third HMFA dimension is characterized by an additional spatial direction on the positive side which includes attributes relating to the perception of space, i.e., several occurrences of the attributes *Echo* and *Reverberation* for the IV experiment and the attributes *Space* and *Amount of echo* for the CV experiment. Note finally that the fourth dimension was difficult to interpret due to the low correlation most attributes with this component.

'Simulated IV data' analysis

To complete this comparative study, the result of an HMFA model using a different hierarchical structure is briefly discussed. As mentioned above, this alternative approach considers the CV data as a set of individual sensory profiles, which relaxes the constraint on the consensus between assessors and allows therefore to test the validity of the panel agreement.

The HMFA model output was explored with the same visualization tools as above

and the main difference to be noted from this investigation relates to the partial score plots. While the global sensory map remained similar to the previous HMFA model, the sensory map of the CV experiment showed a better discrimination and appeared to be more similar to the IV experiment than in the previous HMFA analysis. The fact that an addition of flexibility in CV data modeling[2] improved sample discrimination is an interesting outcome that can be interpreted as a lack of consensus in the CV data set.

10.2.4 Summary

In this section, a combined analysis of the two sensory experiments made on the spatial enhancement systems was reported. The HMFA procedure allowed for a quantitative evaluation of the two sensory maps. The results indicate that the IV data set discriminated better the systems than the CV data set as illustrated by the partial score plots presented above. The study of the Lg plots also revealed that the sensory map resulting from the IV experiment was more complex in a multivariate sense and more consistent across music clips.

The application of HMFA presented in this section demonstrated the utility of this analysis technique. In comparison to the PARAFAC2 technique exploited in Chapter 9, HMFA appears to be more flexible since it only imposes one common mode for the different sets of variables, that is, the spatial enhancement systems in the present case. However, the interpretation of the loadings is more involved than with a PARAFAC2 model due to the large number of model output HMFA provides.

10.3 Comparison of the CV and IV methodologies

To complete this twofold investigation on verbal descriptive analysis techniques, a comparison of the CV and IV methodologies is proposed in this section. Based on the thorough literature review of these two vocabulary development approaches presented earlier and the practical implementation of several experiments of each type by the author, a list of comparative characteristics is provided for these two approaches to highlight their benefits, challenges and limitations and clarify their suitability for different sensory evaluation projects.

10.3.1 Comparative overview of the two methodologies

Table 10.1 gives a comparative overview of the CV and IV methodologies in terms of scope, implementation and outcome. Although the two approaches can be applied to the same set of stimuli as illustrated in the previous section, their scope vary significantly. The sensory characterization has the same verbal descriptive nature in the CV and IV cases but it can be defined as a formal vocabulary validated at a panel level in the former case and only as a global sensory mapping in the latter case. The stimulus selection can be said to be more constrained, at least for rapid IV methods

[2]The data transformations applied to individual sensory profiles by HMFA follow the principles of MFA discussed in Section 7.3.2 (Chapter 7).

Table 10.1: Comparative summary of the CV and IV approaches.

Comparative aspects		**Consensus vocabulary**	**Individual vocabulary**
Scope	Sensory charaterization	- Quantitative descriptive at a panel level - Validated during the vocabulary develpment	- Quantitative descriptive at an individual level - Validated at a panel level after the vocabulary development
	Stimuli	Relatively flexible in selection and size	Well defined and limited in size (at least for the FP and IVP procedures)
	Project type	- High involvement; long-term	- Low involvement and short-term
Implem-entation	Time	Slow: 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 re-used 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

such as FP and IVP, while a larger effort is required for a CV experiment. Overall, it can be said that these two methodologies correspond to very different type of projects. While a IV experiment can be applied for a one-off experiment either for exploring a certain sensory space or evaluating a specific stimulus set, running a CV experiment is more involved but the outcome that can be exploited better on a long-term.

Regarding the practical implementation aspects, large differences can also be noted between the CV and IV methodologies. The time required for such sensory analysis projects is much longer in the case of a CV experiment, e.g. 20 hours vs. 6 hours for the studies reported in this thesis. The type of assessors to be employed is more flexible for the IV method as illustrated by the IVP procedure presented in Chapter 8. The CV development procedure is more complex to implement because it requires group discussions with a panel leader, which is not the case for an IV development. Interestingly, the IV method is favorable in terms of experimental unbias because each assessor is free to develop his/her own set of sensory descriptors without any interaction with other assessors or the experimenter.

The experimental outcome differs also considerably between these two methodologies. The most important difference is the format of the sensory characterization. The large number of individual vocabularies produced in an IV experiment can be considered a disadvantage in comparison to the well defined set of sensory descriptors produced in a CV experiment although the richness of the individual descriptions in the former case can be viewed as a benefit for the exploration of a perceptual domain. The possibility to reuse the developed sensory characterization is also a contrasting factor which should in theory determine the type of methodology to be applied in a sensory project. In terms on analysis and interpretation of results finally, the data treatment is clearly the challenging aspect of the IV methodology as discussed thoroughly in Chapter 7.

10.3.2 Discussion on the two methodologies

At this point, several elements of these two methodologies are briefly discussed. Starting with the CV approach, the concept of 'consensus' and the associated process of 'concept alignment' discussed in Chapter 4 is presented in the sensory science literature as a crucial step in a vocabulary development (O'Mahony, 1991; Munoz and Civille, 1998; Murray *et al.*, 2001). In practice however, this aspect was found to be difficult to implement by the author and the effort required for this process remains unclear as well as the expected outcome in terms of vocabulary accuracy.

Regarding the IV methodology, the author found the multivariate analysis techniques employed for the treatment of individual vocabularies rather coarse in the sense that they do not make any assumption about the semantic meaning of individual attributes but simply look for co-variations in data configurations. Although this type of data treatment is usually successful and can be tested statistically, approaches incorporating semantic information would be valuable. In this respect, the technique developed by Gaines and Shaw (1993) to compare the conceptual systems elicited by experts is interesting because it handles 'distinctions' made between entities and the 'terms' used for these distinctions in a more balanced manner.

Several issues remain open for the author after this investigation on verbal descriptive analysis methods regarding the 'potential' of the quantitative sensory description obtained with the CV and IV methodologies. Especially, the maximum level of accuracy and precision that can be expected with an IV panel and the factors influencing it, e.g., the number of assessors or their experience in sensory analysis, are still unknown.

To conclude on this comparative study, the author would contrast the two methodologies by stating that a well conducted IV experiment should in theory give an equivalent or superior outcome to a CV experiment implemented in a limited time. This implies however that enough assessors are employed (i.e., 15 to 20 subjects) and that a proper elicitation process is implemented including a formal definition of descriptors by the assessors.

Nevertheless, it is clear that IV cannot compete with a thorough CV experiment. In fact, the author believes that the two approaches are complementary and can be combined. For example, an IV experiment can be run as a preliminary step before a more thorough CV experiment because this method allows each assessor to give his or her view, it generates a rich terminology with an unbiased measure of the importance of perceptual aspects and it provides some quantitative results early in the sensory project.

10.4 Conclusion

A comparative analysis of the CV and IV approaches was presented in this chapter. The combined analysis of the data sets resulting from the CV and IV experiments with the HMFA technique highlighted global similarities in the sensory characterization of the spatial enhancement systems but it also showed differences in system discrimination in favor of the IV experiment. The more general comparison of the two methodologies highlighted their respective advantages, limitations and suitability for different sensory analysis projects.

Chapter 11

Summary and conclusions

This thesis reports an investigation on perceptual evaluation methods with a practical application to spatial sound reproduction over headphones.

An overview of perceived quality evaluation is presented in which the term 'Quality' is defined as a measure of the distance between the character of an entity under study and the character of a target associated to this entity. Based on this definition, a novel structured framework for perceived (sound) quality evaluation is proposed making a clear link between the hedonic and sensory characterizations of sound and putting a large emphasis on the sensory domain.

A comprehensive overview of verbal descriptive analysis is provided. This perceptual evaluation approach is positioned in a broader classification of sensory analysis methods first. Then, a thorough review of the concepts, techniques and implementation aspects of the two methodologies available in this category is presented, that is, the consensus vocabulary approach using a panel of assessors to develop a common set of sensory descriptors (Chapter 4) and the individual vocabulary approach letting each assessor of the panel develop his or her own set of descriptors (Chapter 7).

Following the review of the consensus vocabulary approach, the development of a consensus vocabulary of headphone sound perception is reported. A panel of screened assessors was presented with a wide range of headphone sound stimuli and created a consensual set of sixteen attributes with their associated definition, word anchors and audio exemplars to describe this perceptual domain. This vocabulary covers aspects relating to the localization and externalization of sound and the perception of space as well as aspects relating to timbre, loudness and artifacts. It shares some similarities with previous studies on spatial sound perception but represents the first terminology developed specifically for headphone sound to the author's knowledge. The applicability of this vocabulary is demonstrated on a simple set of headphone sound samples and several issues with this type of methodology are discussed such as the anchoring of attribute scales and the agreement of assessors. A listening experiment is finally reported in which a set of eight spatial enhancement systems for headphone reproduction applied to three music clips is evaluated with this consensus vocabulary.

Additionally, the thesis presents a procedure entitled 'individual vocabulary profiling' developed for rapid sensory evaluation of audio applications by inexperienced

assessors and using an individual elicitation approach. IVP is well suited for the unsupervised administration of sensory tests and its modular structure offers some flexibility for the design of experiments tailored for different types of sensory panels. The application of this procedure is illustrated on the set of spatial enhancement systems already evaluated with the consensus vocabulary.

Based on the thorough literature review and the practical implementation of two sets of experiments, a comparison of the consensus vocabulary and individual vocabulary approaches is provided highlighting the benefits, challenges and limitations of each of these two methodologies and clarifying their suitability for different sensory evaluation projects.

The research work reported in this thesis covers also the analysis of data resulting from sensory experiments employing a consensus or an individual vocabulary. This type of experiment generates a large amount of data which can be treated with various analysis techniques. Such tools can help handling and exploring effectively sensory data at a univariate and a multivariate level and are instrumental in the assessment of aspects such as product discrimination level, attribute suitability and assessor or panel performance.

In this thesis, a systematic investigation of the attribute rating data resulting from the two sensory profiling experiments on spatial enhancement systems is reported. Chapter 9 provides a detailed illustration of individual vocabulary data analysis and highlights the importance of multivariate data analysis in this context and the challenges of interpreting semantically the underlying sensory dimensions.

Several advanced multivariate data analysis methods are explored in this work. Multiple factor analysis (MFA) and hierarchical multiple factor analysis (HMFA) are applied for the combined analysis of several sets of multivariate data (Chapters 6 and 9 respectively). A novel application of the PARAFAC2 multi-way data analysis technique is presented for handling the four-way data structure of the individual vocabulary data relating to spatial enhancement systems. One advantage of these analysis techniques is to provide a quantitative description of the spatial enhancement algorithms at a global level while highlighting music clip differences.

Through these different analyses, the main perceptual differences between spatial enhancement systems for headphone reproduction are identified and quantified. In addition, the comparative analysis applied to the results of the two experiments illustrates a slightly better sample discrimination with the IV experiment and an easier semantic interpretation with the CV experiment.

The aspects covered in this thesis concern only one side of the quality evaluation framework described in Chapter 3. Beyond the sensory evaluation domain, the preference mapping matter forms another challenging topic not covered in this thesis. This field of sensory science is actively researched and would deserve more attention for testing the validity extent of the perceptual quality evaluation framework proposed in this work.

Appendix A

Attribute graphs per music clip (Chapter 6)

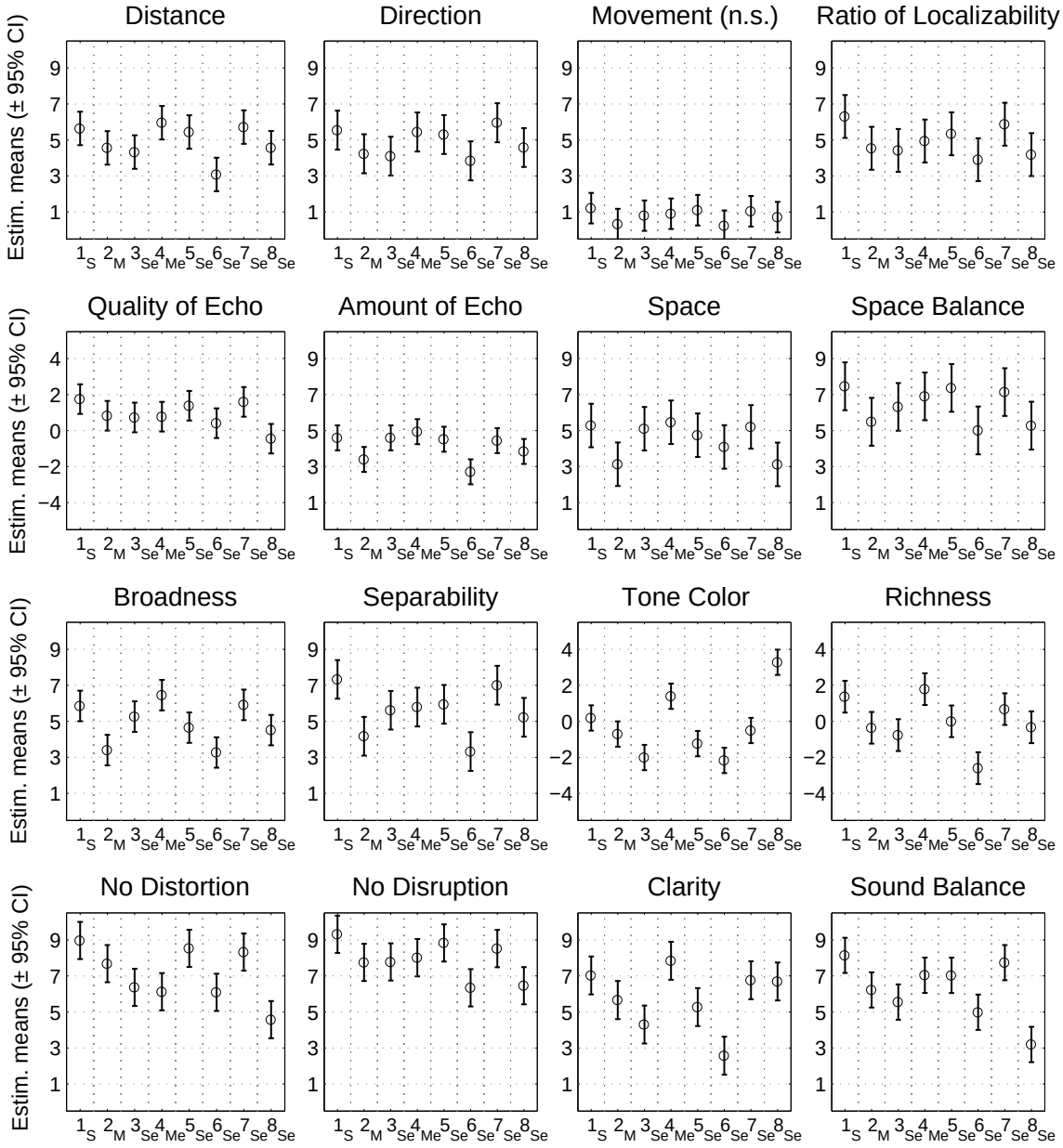


Figure A.1: Mean scores estimated from ANOVA and 95% CI of the eight spatial enhancement systems on 16 attributes for the music clip #2.

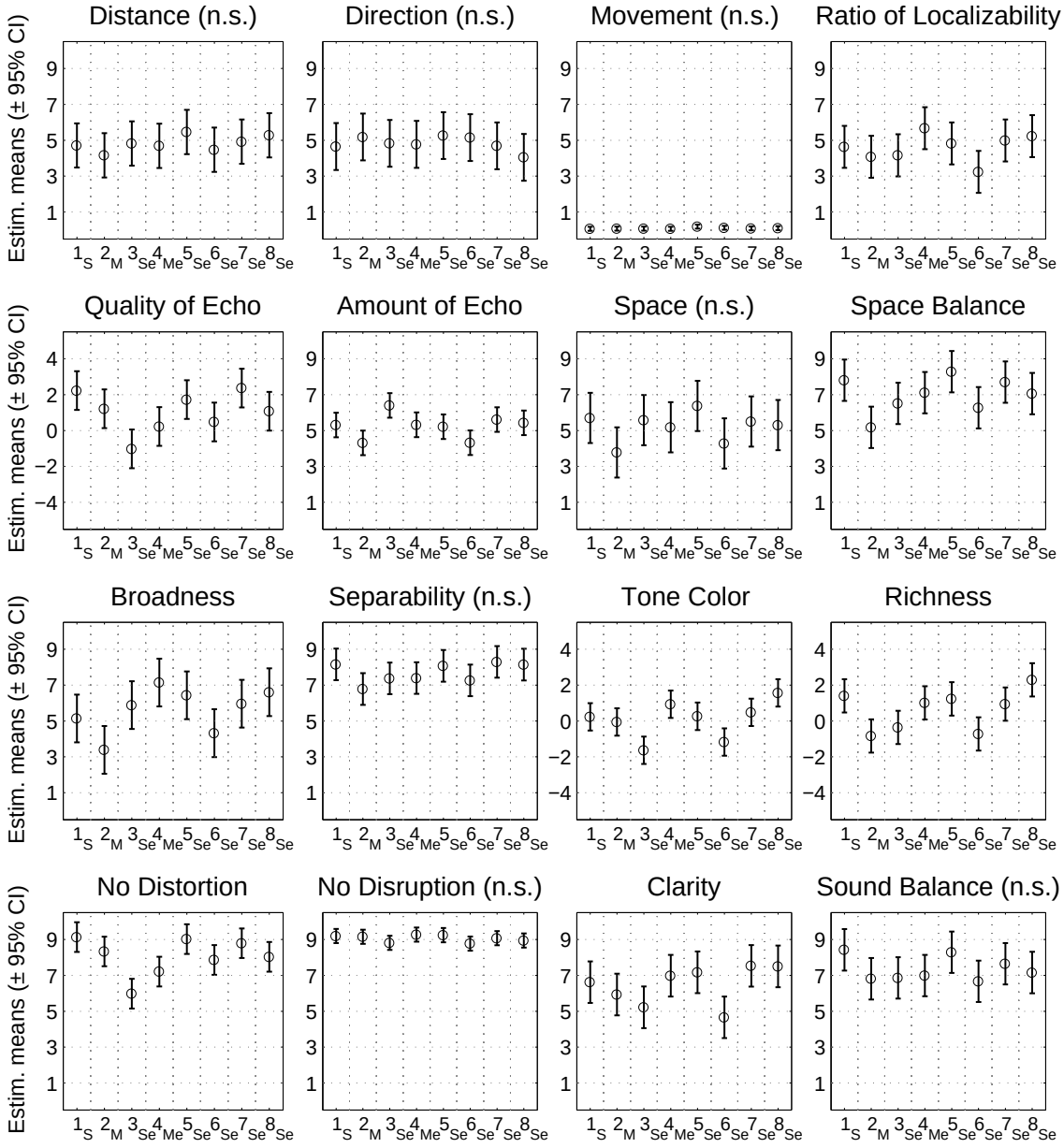
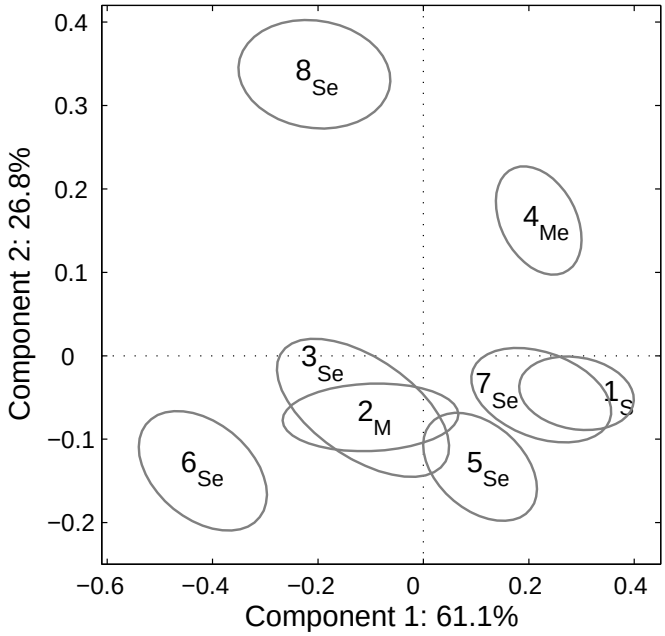


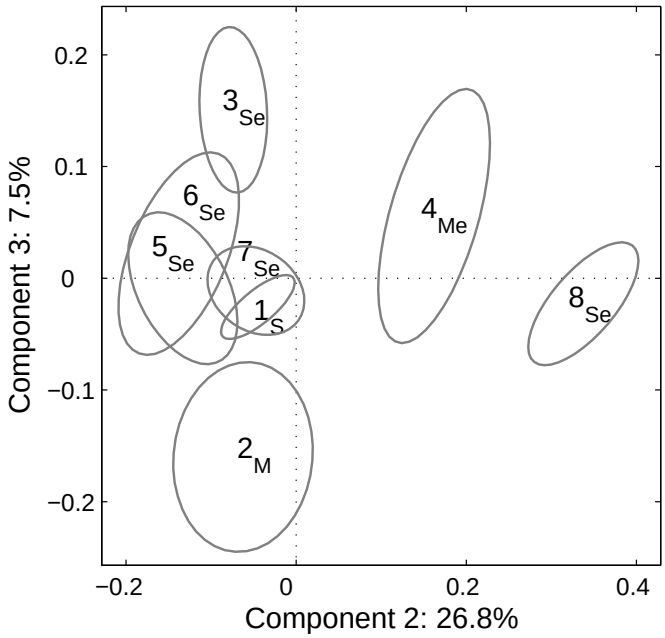
Figure A.2: Mean scores estimated from ANOVA and 95% CI of the eight spatial enhancement systems on 16 attributes for the music clip #3.

Appendix B

Illustration of PCA results per music clip (Chapter 6)

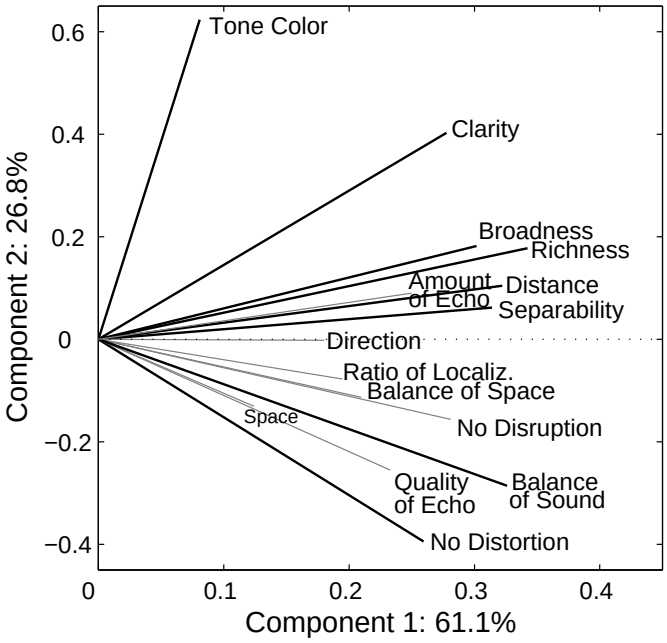


(a) PCA scores - PCs 1–2

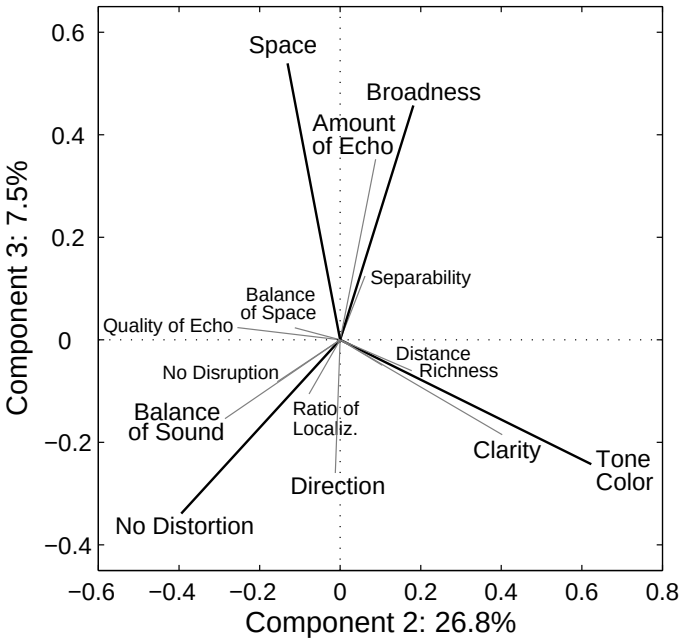


(b) PCA scores - PCs 2–3

Figure B.1: Three first components of a PCA applied to the music clip #2. Ellipses around the PCA scores represent 95% confidence regions based on assessor variability.

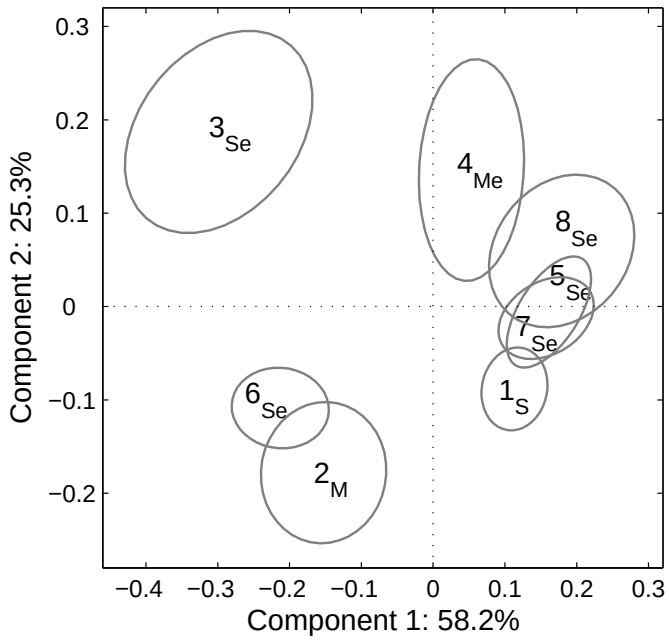


(a) PCA loadings - PCs 1–2

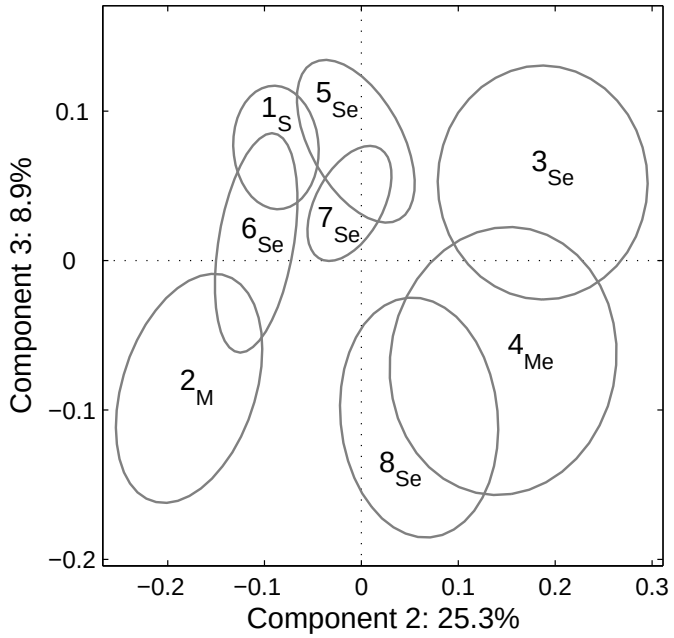


(b) PCA loadings - PCs 2–3

Figure B.2: Three first components of a PCA applied to the music clip #2. These loading plots illustrate the relationship between the attributes and the PC dimensions.

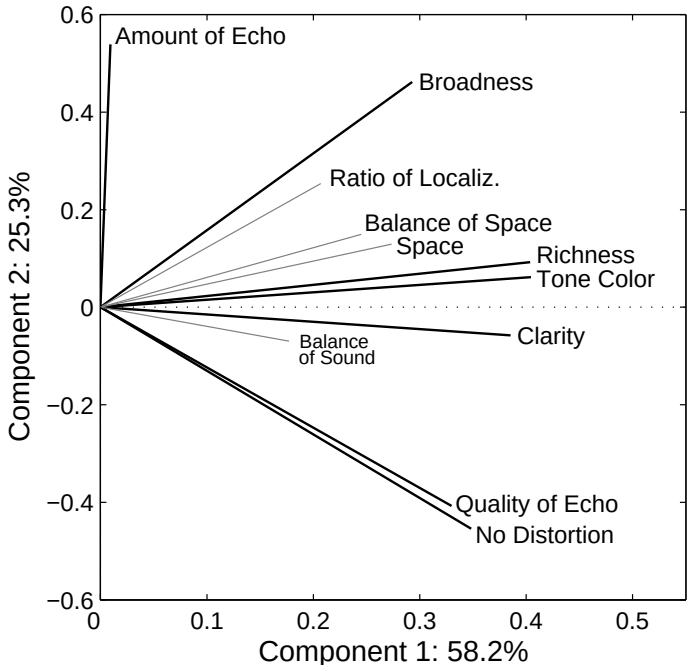


(a) PCA scores - PCs 1–2

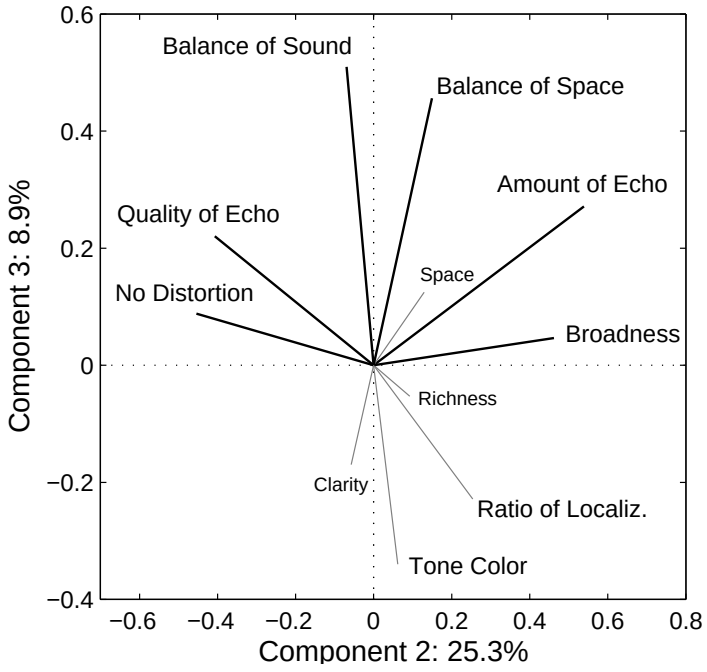


(b) PCA scores - PCs 2–3

Figure B.3: Three first components of a PCA applied to the music clip #3. Ellipses around the PCA scores represent 95% confidence regions based on assessor variability.



(a) PCA loadings - PCs 1–2



(b) PCA loadings - PCs 2–3

Figure B.4: Three first components of a PCA applied to the music clip #3. These loading plots illustrate the relationship between the attributes and the PC dimensions.

Appendix C

Matlab implementation of the generalized Procrustes analysis procedure (Chapter 7)

```
function [Xgpa,Ggpa,Q,isoscales] = gpa(X);
% GPA.M minimizes the generalized (orthogonal) Procrustes criteria for a set of configurations
%       with an equal number of objects
% - Input:
%      X: 'K-sets' array: Cell array comprising of several configurations,
%                         i.e. 2-way matrices of size Objects x Variables.
%                         Each 2-way matrix need to have the same number of objects
%                         The number of variables can vary from one configuration to another, e.g.
%                              X{1} = 8 Products x 7 Attributes for Assessor 1,
%                              X{2} = 8 Products x 5 Attributes for Assessor 2,
%                              ...,
%                              X{n} = 8 Products x 9 Attributes for Assessor n.

% - Output:
%       Xgpa: 3-way array of GPA transformed individual configurations
%             size: Objects x max(Variables) x Configurations
%             The second mode represents the common dimensions derived by GPA
%       G: 2-array matrix containing the 'group average' configuration
%          size: Objects x max(Variables)
%       Q: 3-array matrix of individual rotation matrices
%       isoscales: vector containing the isotropic scale of each individual configuration
%
% - Example: for i = 1:15, X{i} = 10*rand([12 ceil(10*rand)]); end;
%            [Xgpa,G,Q,isoscales] = gpa(X);
%
% This routine uses isoscaling.m and procrust.m (attached below)
% Gaetan Lorho
% 2004

%% Step 0 - Check input
Co = size(X,2);
for c = 1:Co, if find(isnan(X{c})),
    error(['Missing values present in dataset (configuration # ', num2str(c),...
        ' can not be handled with this version of gpa.m']), end, end
for c = 1:Co, allobj(c) = size(X{c},1); end
if allobj - mean(allobj), error(['The number of objects differs across configurations ',...
        'or the format of the input cell array is wrong']); end
for c = 1:Co, allvar(c) = size(X{c},2); end

%% Step 1 - Centering +  Step 2 - Zero padding
%  this gives 3-way matrix of size Objects x max(Variables) x Configurations
Xc = repmat(0,[allobj(1) max(allvar) Co]);
for c = 1:Co,
    Xc(1:allobj(c),1:allvar(c),c) = X{c} - ones(size(X{c},1),1)*mean(X{c});
```

```
end
[Ob,Va,Co] = size(Xc);
ftext = [sprintf('\n'),'Dataset contains ', num2str(Ob), ' objects and ', num2str(Co),...
    ' configurations', sprintf('\n\n'),'Size of individual configurations :',sprintf('\n\n')];
for c = 1:Co,
    ftext = [ftext, 'Configuration ', num2str(c), ' : ', num2str(allvar(c)), ' variable(s)',...
        sprintf('\n')];
end; disp(ftext);

%% Step 3 - Apply scaling & rotation/reflection until convergence
% Method based on Gower & Dijksterhuis (2004), Procrustes Problems, Oxford University Press
% Algorithm 9.3 (page 114) is applied using an orthogonal transformation matrix (Tk = Qk)

%% Step 3.1 - Initialisation
Gc = mean(Xc,3); % centered 'group average' configuration
SSRc = 0; for c = 1:Co,
    R = Gc - Xc(:,:,c); SSRc = trace(R'*R) + SSRc; end  % centered residual sum of squares
ConvergenceStep = SSRc;  % initial convergence step
for c = 1:Co,
    Q(:,:,c) = procrust(squeeze(Xc(:,:,c)),squeeze(mean(Xc(:,:,[1:c-1 c+1:end]),3)));
    % Orthogonal rotation of the original (centered) individual configuration 'c'
    % to fit the initial 'c-excluded' group average
end

%% Step 3.2 - Iterative scaling / Procrustes transformations
iteration = 1; disp(['Iterations performed before algorithm convergence : ',sprintf('\n')] );
while (ConvergenceStep > 0.0001) & iteration < 1000,
    disp(['Iteration #',num2str(iteration),' - Residual sum of squares : ',num2str(SSRc)] );
    iteration = iteration+1;
        for c1 = 1:Co, for c2 = 1:Co,  % update S matrix (Sij = trace(Qi'Xi'XjQj))
                S(c1,c2) = trace(squeeze(Q(:,:,c1))'*squeeze(Xc(:,:,c1))'...
                    *squeeze(Xc(:,:,c2))*squeeze(Q(:,:,c2)));
        end, end;
    isoscales = isoscaling(S);  % estimate isotropic scales based on current rotated configurations
    for c = 1:Co,  % update transformed individual configurations
        Xgpa(:,:,c) = isoscales(c)*squeeze(Xc(:,:,c))*squeeze(Q(:,:,c));
    end;
    for c = 1:Co,  % apply Procrustes transformation to current transformed individual configurations
        Q(:,:,c) = procrust(isoscales(c)*squeeze(Xc(:,:,c)),squeeze(mean(Xgpa(:,:,[1:c-1 c+1:end]),3)));
        % Orthogonal rotation of scaled original c individual configuration
        %     to fit the current c-excluded group average
    end;
    % test for satisfactory convergence
    for c = 1:Co,  % current transformed individual configurations
        Xgpa(:,:,c) = isoscales(c)*squeeze(Xc(:,:,c))*squeeze(Q(:,:,c));
    end;
    Ggpa = mean(Xgpa,3);  % current 'group average' configuration
    SSR = 0; for c = 1:Co,  % current residual sum of squares
        R = Ggpa - Xgpa(:,:,c); SSR = trace(R'*R) + SSR;
    end;
    ConvergenceStep = SSRc - SSR; SSRc = SSR;  % updated convergence step and residual sum of squares
end;
if iteration == 1000,
    disp([sprintf('\n'),'Caution! The algorithm did not converge...',sprintf('\n')] );
end;


function iscale = isoscaling(S);
% Isotropic scaling in generalized Procrustes analysis
% S is a matrix containing the traces of the product of the different individual configurations,
% that is, Sij = trace(Xi'Xj) with 1<i<Co and 1<j<Co (size(S) = assessors x assessors)
% This algorithm derives the scaling constants iscale(1,...,Co) maximizing 'h' defined as
%     h(1,...,Co) = Sum [iscale(i)*iscale(j) * trace(Xi'Xj)] under the constraint
%     Sum [iscale(i)^2 * trace(Xi'Xi)] = Sum [trace(Xi'Xi)] = m.
% Reference: ten Berge (1977), Orthogonal Procrustes Rotation for 2 or more Matrices,
%     Psychometrika, 42, no.2, 1977.
Phi = (diag(diag(S))^(-0.5))*S*(diag(diag(S))^(-0.5));  % Matrix of coefficients of congruence
[eigvec,eigval] = eig(Phi + eps);  % eigenvalue-eigenvector decomposition
[y,x] = max(diag(eigval));   % The eigenvector associated with the largest eigenvalue
```

```
p1 = abs(eigvec(:,x));       % contains the scaling factors of interest
for i = 1:size(S,1),
    iscale(i) = sqrt(sum(diag(S))/S(i,i))*p1(i); % Compute isotropic scaling for each configuration
end

function Q = procrust(X,Y);
% Orthogonal Procrustes problem: find the orthogonal rotation giving the best match between X and Y
% The least-squares solution is to minimize ||Y - X*Q||,
%     which is equivalent to maximizing trace(Q'X'Y)
% The maximizing Q can be found by calculating [U,D,V] = svd(X'*Y)
%     and the solution matrix is Q = U*V';
% Reference: Chapter 4 of Gower & Dijksterhuis (2004), Procrustes Problems, Oxford University Press
[U,D,V] = svd(X'*Y + eps);
Q = U*V';
```

Appendix D

Illustration of GPA-PCA results per music clip (Chapter 9)

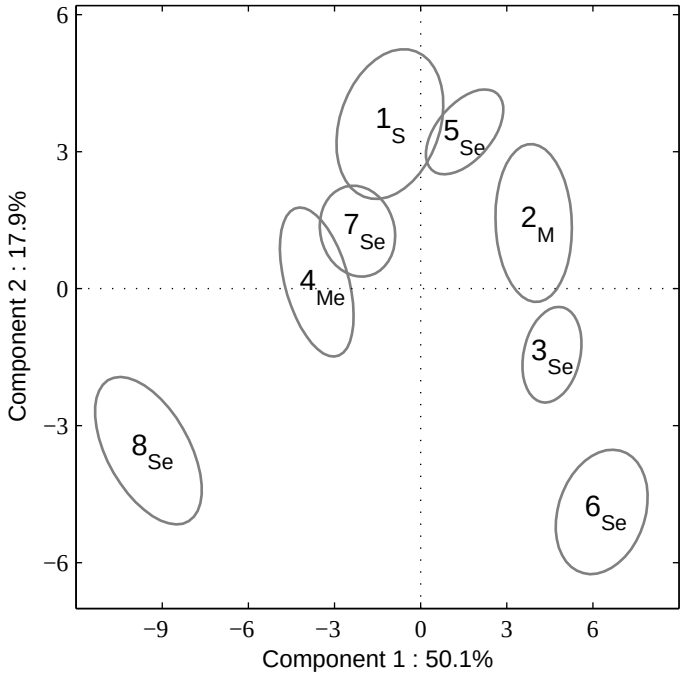
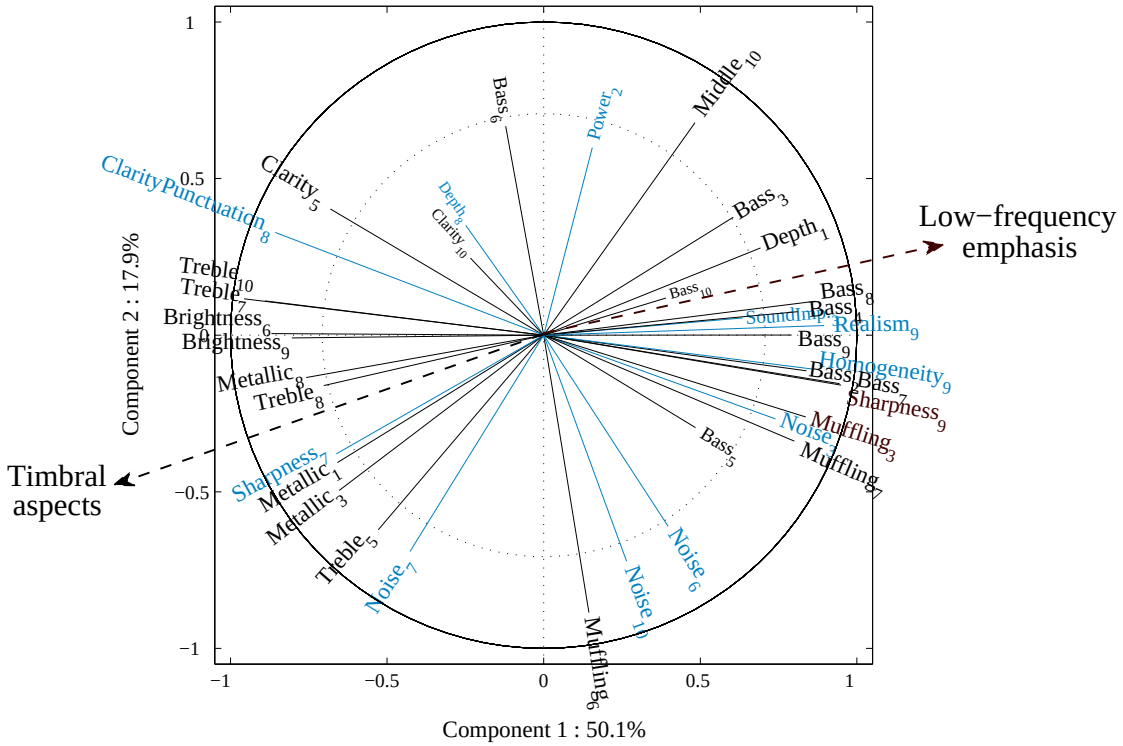
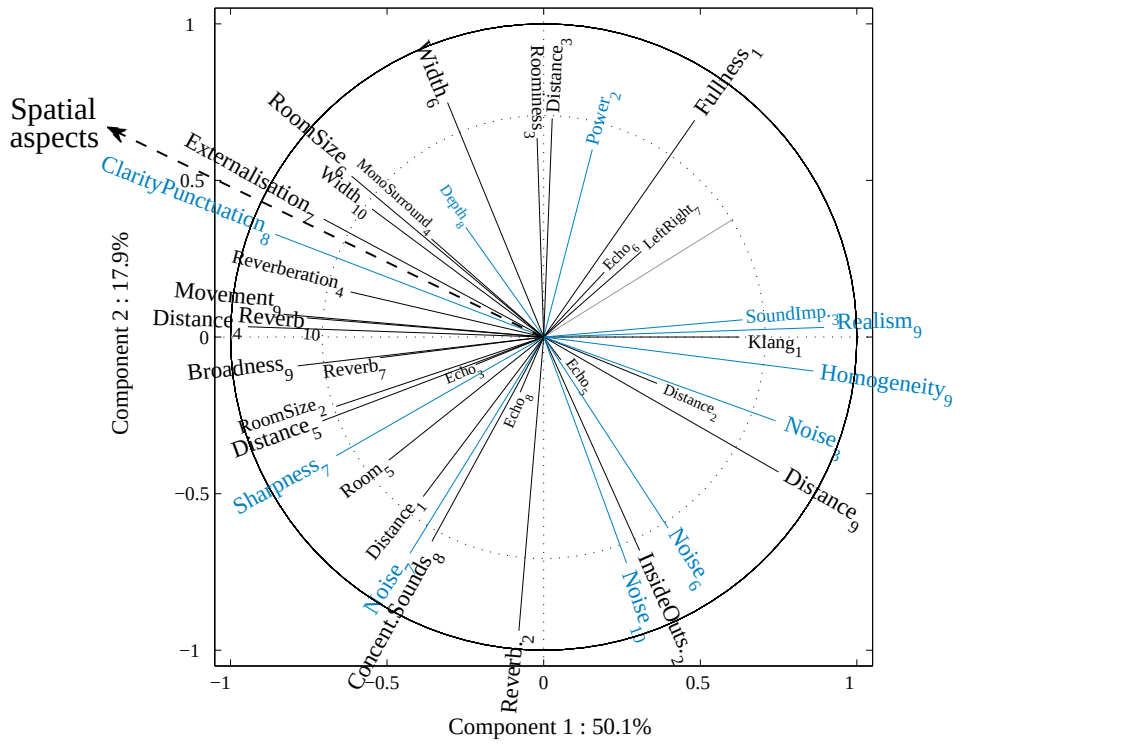


Figure D.1: PCA analysis of the GPA average configuration obtained for the music clip #2 – Score plot.



(a) Timbral aspects (refine low-freq arrow)



(b) Spatial aspects

Figure D.2: PCA analysis of the GPA average configuration obtained for the music clip #2 – Correlation loading plots.

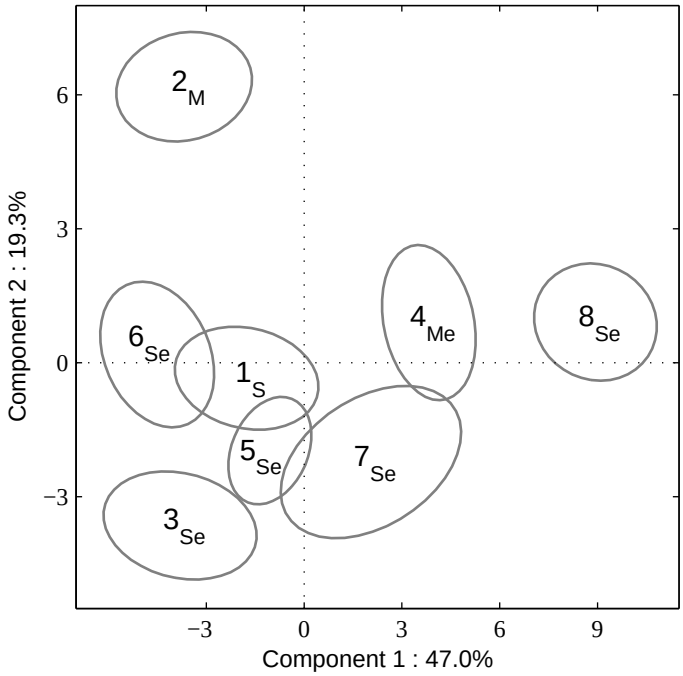
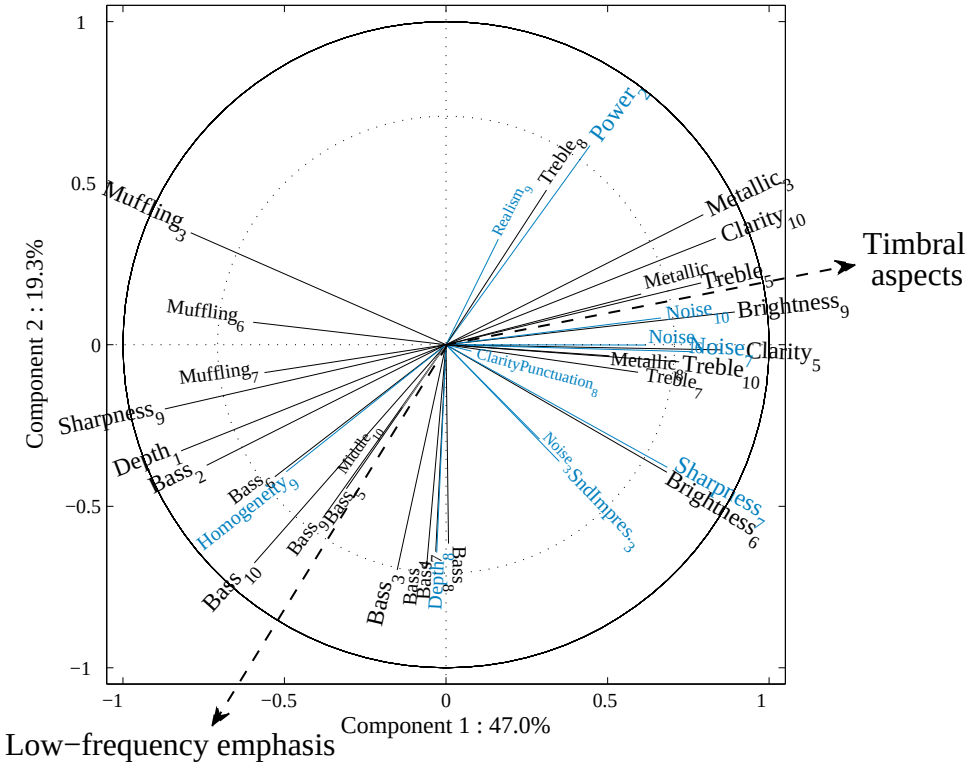
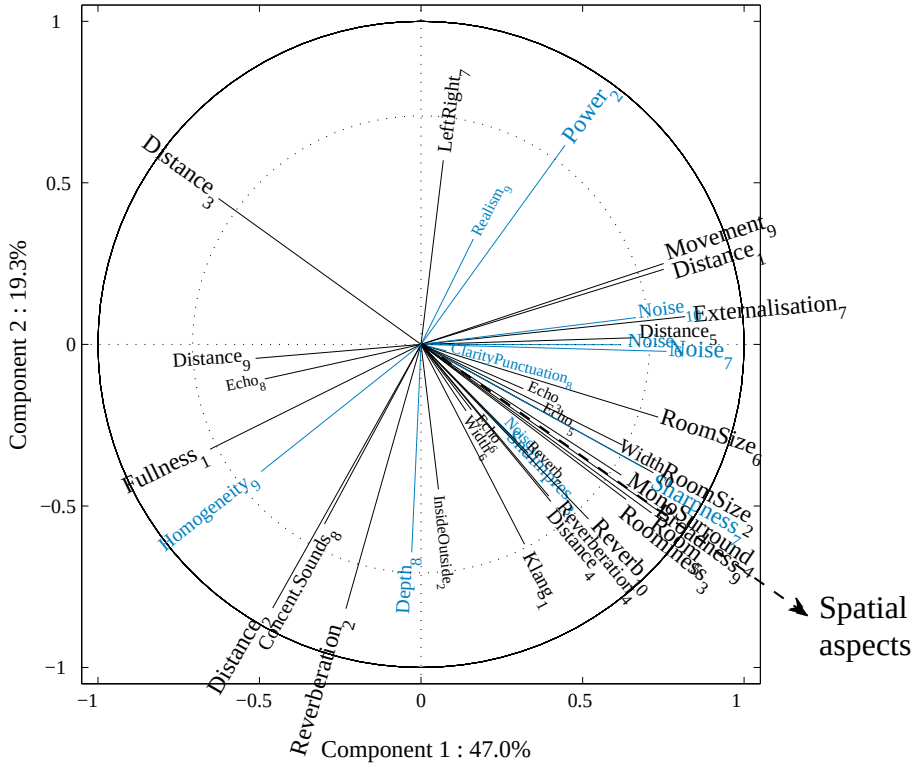


Figure D.3: PCA analysis of the GPA average configuration obtained for the music clip #3 – Score plot.



(a) Timbral aspects



(b) Spatial aspects

Figure D.4: PCA analysis of the GPA average configuration obtained for the music clip #3 – Correlation loading plots.

Appendix E

Overview of PARAFAC2 (Chapter 9)

This appendix gives a mathematical overview to the PARAFAC2 model applied to 4-way individual sensory profiling data in Chapter 9. An introduction to PARAFAC1, the original version of the parallel factor analysis model, is provided first for 3-way sensory data (Figure E.1) and the principle is then extended to the PARAFAC2 model for 3-way (Figure E.2) and 4-way individual sensory data.

The PARAFAC model can be expressed in a matrix notation as $\mathbf{X}_k = \mathbf{T}\mathbf{D}_k\mathbf{P}' + \mathbf{R}_k$ $(k = 1, \ldots, K)$ where $\mathbf{X}_k$ is a slab of the 3-way array $\underline{\mathbf{X}}$ and represents the profile of the assessor k in a consensus sensory experiment (see Figure 6.2 in Chapter 6), i.e., $\mathbf{X}_k$ is a two-way matrix of size I systems $\times$ J sensory attributes. In the equation of the L-component model illustrated in Figure E.1, $\mathbf{T}$ is an $I \times L$ matrix of factor scores ($\mathbf{T}$ relates to the I systems in a sensory data set), $\mathbf{P}$ is a $J \times L$ matrix of factor loadings ($\mathbf{P}$ relates to the J attributes), $\mathbf{D}_k$ is a diagonal $L \times L$ matrix containing the weights for the k^{th} slab of $\underline{\mathbf{X}}$ ($\mathbf{D}_k$ relates to the L factor weights[1] of the assessor k), and $\mathbf{R}_k$ denotes an $I \times J$ matrix of residuals. The PARAFAC model is fitted to the array $\underline{\mathbf{X}}$ through a procedure of alternating least squares aiming to minimize the sum of squared residuals in $\underline{\mathbf{R}}$ and has the interesting property of giving unique solutions.

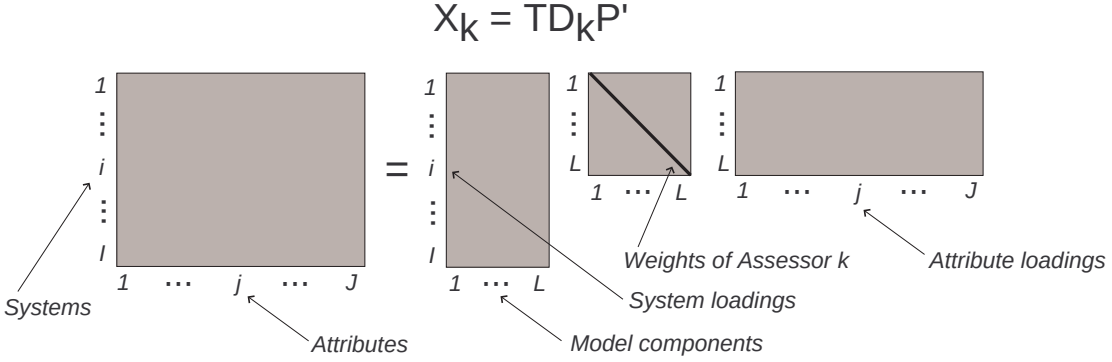


Figure E.1: Two-way representation of the PARAFAC model.

[1]In PARAFAC models, the term 'loading' is often employed without distinction for all the modes.

PARAFAC2 uses a similar decomposition approach but has the ability to handle data sets with a 'free' mode, that is, a mode in which variables can differ in number, might be shifted across another mode or/and are not commensurable. Individual sensory profiles follow this data format as the number of columns and the nature of the attributes in the matrix $\mathbf{X}_k$ can vary across assessors. This idea was introduced by Harshman (1972) who replaced $\mathbf{P}$ in the PARAFAC model equation by a matrix $\mathbf{P}_\mathrm{k}$ that can vary across k to handle matrices $\mathbf{X}_k$ of varying size $I \times J_k$ $(k = 1, \ldots, K)$. Harshman noted that such a model can also give unique solutions if the cross-product matrix $\mathbf{P}_\mathrm{k}'\mathbf{P}_\mathrm{k}$ remains constant over k and he proposed to fit the model indirectly with the help of the cross-product matrix $\mathbf{X}_\mathrm{k}\mathbf{X}_\mathrm{k}'$.

Kiers *et al.* (1999) proposed a direct method to fit the PARAFAC2 model as $\mathbf{X}_k = \mathbf{T}\mathbf{D}_k(\mathbf{Q}_k\mathbf{F})' + \mathbf{R}_k$ (see Figure E.2). In this equation, the matrices $\mathbf{T}$ and $\mathbf{D}_k$ are defined as in the PARAFAC model but $\mathbf{P}$ is replaced by two new matrices, that is, a quadratic matrix $\mathbf{F}$ of size $L \times L$ and a columnwise orthonormal matrix $\mathbf{Q}_k$ of size $J_k \times L$ $(J_k \geq L)$. It can be noted that the cross-product matrix of $\mathbf{Q}_k\mathbf{F}$ remains also constant over k in this model formulation because $(\mathbf{Q}_k\mathbf{F})'(\mathbf{Q}_k\mathbf{F}) = \mathbf{F}'\mathbf{F}$. This direct PARAFAC2 model fitting approach provides model parameters for all three modes and offers several advantages in terms of algorithm implementation. The procedure proposed by Kiers *et al.* (1999) is in fact equivalent to applying (iteratively) an ordinary PARAFAC model to the three-way array $\underline{\mathbf{Y}}$ consisting of the two-way matrices $\mathbf{X}_k\mathbf{Q}_k$ $(k = 1, \ldots, K)$ which all have the same size $I \times L$.

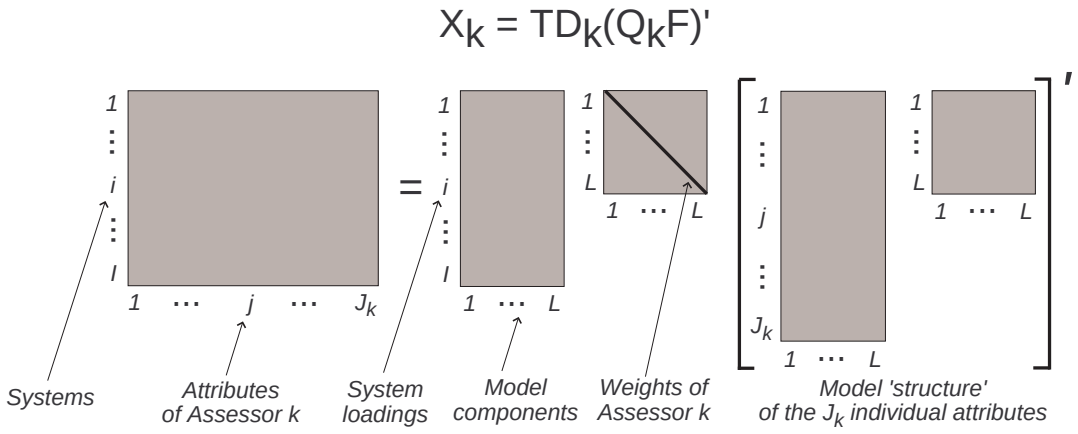


Figure E.2: Two-way representation of the PARAFAC2 model.

Both versions of PARAFAC can be extended to higher orders. For example, to apply the PARAFAC2 model proposed by Kiers *et al.* (1999) to the four-way data set presented in Chapter 9, the two-way array $\mathbf{X}_k$ is replaced by the three-way array $\underline{\mathbf{X}_m}$ of size $I \times J \times K_m$ (i.e., `System` × `Music clip` × `Attribute`) for the assessor $m = 1, \ldots, M$. A similar orthogonal matrix $\mathbf{Q}_m$ is introduced to handle variations of the mode `Attribute` across the mode `Assessor` and following the direct PARAFAC2 model fitting approach described above, the array consisting of the unfolded slabs $\mathbf{X}_m\mathbf{Q}_m$ $(m = 1, \ldots, M)$ can be modeled by an ordinary four-way PARAFAC model.

Bibliography

Algazi, V. R., Avendano, C., and Duda, R. O. (**2001**). "Elevation localization and head-related transfer function analysis at low frequencies," *Journal of the Acoustical Society of America* **109**, 1110–1122.

Anderson, T. W. (**2003**). MANOVA: An Introduction to Multivariate Statistical Analysis, Wiley, Hoboken, NJ, USA.

ANSI S3.5 (**1997**). Methods for Calculation of the Speech Intelligibility Index, American National Standards Institute.

Arnold, G. M. (**1992**). "Scaling factors in generalized procrustes analysis," in I. Y. Dodge and J. Whittaker, eds, *Computational statistics, vol.1. Proceedings of the 10^{th} symposium on computational statistics, COMPSTAT.*

Arnold, G. M. and Williams, A. A. (**1986**), *Statistical procedures in food research*, Elsevier Applied Science, chapter 7: The use of Generalised Procrustes Analysis techniques in sensory analysis, pp. 233–253.

Arnold, G. M., Gower, J. C., Gardner-Lubbe, S., and Roux, N. J. L. (**2007**). "Biplots of free-choice profile data in generalized orthogonal procrustes analysis," *Journal of the Royal Statistics Society, Series C: Applied Statistics* **56**, 445–458.

Asano, F., Suzuki, Y., and Sone, T. (**1990**). "Role of spectral cues in median plane localization," *Journal of the Acoustical Society of America* **88**, 159–168.

Atal, B. S. and Schroeder, M. R. (**1966**). "Apparent sound source translator," U.S. patent no. 3,236,949.

Barron, M. (**1988**). "Subjective study of British symphony concert halls," *Acustica* **66**, 1–14.

Barron, M. F. E. (**1971**). "The subjective effects of first reflections in concert halls – The need for lateral reflections," *Journal of Sound and Vibration* **15**, 475–494.

Bauer, B. B. (**1961**). "Stereophonic earphones and binaural loudspeakers," *Journal of the Audio Engineering Society* **9**, 148–151.

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

Bech, S. (**1992**). "Selection and training of subjects for listening tests on sound-reproducing equipment," *Journal of the Audio Engineering Society* **40**, 590–610.

Bech, S. (**1993**). "Training of subjects for auditory experiments," *Acta Acustica* **1**, 89–99.

Bech, S. (**1994**). "Perception of timbre of reproduced sound in small rooms: Influ-

ence of room and loudspeaker position," *Journal of the Audio Engineering Society* **42**, 999–1007.

Bech, S. and Zacharov, N. (**2006**). Perceptual Audio Evaluation – Theory, Method and Application, John Wiley & Sons, Chichester, England.

Begault, D. R. (**1992**). "Perceptual effects of synthetic reverberation on three-dimensional audio systems," *Journal of the Audio Engineering Society* **40**, 895–904.

Begault, D. R. (**1994**). 3-D Sound for Virtual Reality and Multimedia, Academic Press, Cambridge, MA, USA.

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (**2000**). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *Journal of the Audio Engineering Society* **49**, 904–916.

Beranek, L. L. (**1962**). Music, acoustics and architecture, John Wiley, New York.

Beranek, L. L. (**1996**). Concert and opera halls: How they sound, Acoustical Society of America, Woodbury, NY, USA.

Berg, J. (**2005**). "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.

Berg, J. and Rumsey, F. (**1999a**). "Identification of perceived spatial attributes of recordings by repertory grid technique and other methods," in *Proceedings of the 106th Convention of the Audio Engineering Society*, Munich, Germany.

Berg, J. and Rumsey, F. (**1999b**). "Spatial attribute identification and scaling by repertory grid technique and other methods," in *Proceeding of the 16th International Conference of the Audio Engineering Society*, Rovaniemi, Finland.

Berg, J. and Rumsey, F. (**2000**). "In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors," in *Proceedings of the 108th Convention of the Audio Engineering Society*, Paris, France.

Berkhout, A. J., Vries, D. D., and Vogel, P. (**1993**). "Acoustic control by wave field synthesis," *Journal of the Acoustical Society of America* **93**, 2764–2778.

Bi, J. (**2003**). "Agreement and reliability assessments for performance of sensory descriptive panel," *Journal of Sensory Studies* **18**, 61–76.

Blauert, J. (**1997**). Spatial hearing. The psychophysics of human sound localisation, MIT press, Cambridge, MA, USA.

Blauert, J. and Jekosch, U. (**2003**). "Concepts behind sound quality: Some basic considerations," in *Proceedings of the the 32nd International Congress and Exposition of Noise Control Engineering*, number N466, Inter-Noise, Seogwipo, Korea.

Blumlein, A. D. (**1931**). "Improvements in and relating to sound-transmission, sound-recording and sound-reproduction systems," U.K. patent no. 394,325.

Boone, M. M., Bruijn, W. P. J. D., and Horbach, U. (**1999**). "Virtual surround speakers with wave field synthesis," in *Proceedings of the 106th Convention of the Audio Engineering Society*, Munich, Germany.

Boone, M. M., Verheijen, E. N. G., and Tol, P. F. V. (**1995**). "Spatial sound-field

reproduction by wave-field synthesis," *Journal of the Audio Engineering Society* **43**, 1003–1012.

Brandt, M. A., Skinner, E. Z., and Coleman, J. A. (**1963**). "Texture profile method," *Journal of Food Science* **28**, 404–409.

Bro, R. and Smilde, A. K. (**2003**). "Centering and scaling in component analysis," *Journal of Chemometrics* **17**, 16–33.

Bro, R., Qannari, E. M., Kiers, H. A. L., Naes, T., and Frost, M. B. (**2008**). "Multiway models for sensory profiling data," *Journal of Chemometrics* **22**, 36–45.

Brockhoff, P. B. (**2001**). "Sensory profile average data: Combining mixed model ANOVA with measurement error methodology," *Food Quality and Preference* **12**, 413–426.

Brockhoff, P. M. (**1998**). "Assessor modeling," *Food Quality and Preference* **9**, 87–89.

Brockhoff, P. M. (**2003**). "Statistical testing of individual differences in sensory profiling," *Food Quality and Preference* **14**, 425–434.

Brockhoff, P. M. and Skovgaard, I. M. (**1994**). "Modelling individual differences between assessors in sensory evaluations," *Food Quality and Preference* **5**, 215–224.

Brockhoff, P. M., Hirst, D., and Næs, T. (**1996**), *Multivariate analysis of data in sensory science*, Vol. 16 of *Data handling in science and technology* Næs and Risvik (1996), chapter 10: Analysing individual profiles by three-way factor analysis, pp. 307–342.

Brown, C. P. and Duda, R. O. (**1998**). "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio processing* **6**, 476–488.

Cabrera, D., Ferguson, S., and Schubert, E. (**2007**). "Psysound3: Software for acoustical and psychoacoustical analysis of sound recordings," in *Proceedings of the International Conference on Auditory Display*, ICAD, Montréal, Canada.

Cairncross, W. E. and Sjöström, L. B. (**1950**). "Flavor profile – a new approach to flavor problems," *Food Technology* **4**, 308–311.

Carlile, S. (**1996**). Virtual Auditory Space: Generation and Applications, Chapell & Hall, Austin.

Carroll, J. D. (**1972**). "Individual differences and multidimensional scaling," *Multidimensional Scaling; Theory and Applications in the Behavioral Sciences* pp. 105–155.

Carroll, J. D. and Chang, J. J. (**1970**). "Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition," *Psychometrika* **35**, 283–319.

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

Caul, J. (**1957**). "The profile method of flavor analysis," *Advances in Food Research* **7**, 1–40.

CCITT (**1992**). Handbook of telephonometry, International Telecommunications Union.

Chang, J. J. and Carroll, J. D. (**1969**). "How to use MDPREF, a computer program for multidimensional analysis of preference data," *Computer manual, Murray Hill,*

NJ: Bell Telephone Laboratories.

Choisel, S. (**2003**). "Pointing technique with visual feedback for sound source localization experiments," in *Proceedings of the 115th Convention of the Audio Engineering Society*, New York, NY, USA.

Choisel, S. and Wickelmaier, F. (**2005**). "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," in *Proceedings of the 118th Convention of the Audio Engineering Society*, Barcelona, Spain.

Choisel, S. and Wickelmaier, F. (**2006**). "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," *Journal of the Audio Engineering Society* **54**, 815.

Choisel, S. and Wickelmaier, F. (**2007**). "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *Journal of the Acoustical Society of America* **121**, 388.

Chung, W., Carlile, S., and Leong, P. (**2000**). "A performance adequate computational model for auditory localization," *Journal of the Acoustical Society of America* **107**, 432.

Civille, G. C. (**1979**), *Sensory evaluation methods for the practicing food technologist*, Institute of Food Technologists, M. R. Johnson, Ed., Chicago, Illinois, USA, chapter 6: Descriptive analysis.

Civille, G. V. and Lawless, H. T. (**1986**). "The importance of language in describing perception," *Journal of Sensory Studies* **1**, 203–215.

Couronne, T. (**1997**). "A study of assessors' performance using graphical methods," *Food Quality and Preference* **8**, 359–365.

Dahl, T., Tomic, O., Wold, J. P., and Næs, T. (**2008**). "Some new tools for visualising multi-way sensory data," *Food Quality and Preference* **19**, 103–113.

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

Dairou, V., Sieffermann, J. M., Priez, A., and Danzart, M. (**2003**). "Sensory evaluation of car brake systems. The use of Flash profile as a preliminary study before a conventional profile," in *SAE World Congress*, Detroit, MI.

Danzart, M., Sieffermann, J. M., and Delarue, J. (**2004**). "New developments in preference mapping techniques, finding out a consumer optimal product, its sensory profile and the key sensory attributes," in *7th Sensometrics Meeting*, Davis, CA, USA.

de Jong, S., Heidema, J., and van der Knaap, H. C. M. (**1998**). "Generalized Procrustes analysis of coffee brands tested by five European sensory panels," *Food Quality and Preference* **9**, 111–114.

Delarue, J. and Sieffermann, J. M. (**2004**). "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**, 383–392.

Dijksterhuis, G. (**1995**). "Assessing panel consonance," *Food Quality and Preference* **6**, 7–14.

Dijksterhuis, G. (**1996**), *Multivariate analysis of data in sensory science*, Vol. 16 of *Data handling in science and technology* Næs and Risvik (1996), chapter 7: Procrustes analysis in sensory research, pp. 185–217.

Dijksterhuis, G. B. and Gower, J. C. (**1991**). "The interpretation of generalized procrustes analysis and allied methods," *Food Quality and Preference* **3**, 67–87.

Dijksterhuis, G. B. and Heiser, W. J. (**1995**). "The role of permutation tests in exploratory multivariate data analysis," *Food quality and preference* **6**, 263–270.

Dray, S. (**2008**). "On the number of principal components: A test of dimensionality based on measurements of similarity between matrices," *Computational Statistics and Data Analysis* **52**, 2228–2237.

Du Moncel, T. (**1887**), *Le téléphone de M. Ader*, 5^{th} edn, Librairie Hachette, Paris, France, pp. 117–127.

Duda, R. O. and Martens, W. L. (**1998**). "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America* **104**, 3048–3058.

Durlach, N. I., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., and Wenzel, E. M. (**1992**). "On the externalization of auditory images," *Presence: Teleoperators and Virtual Environments* **1**, 251–257.

EBU 3276 (**1998**). Technical document Tech 3276 – Listening conditions for the assessment of sound programme material: monophonic and two-channel stereophonic, 2^{nd} edn, European Broadcast Union.

EBU 3276-1 (**1999**). Technical document Tech 3276: Supplement 1 – Listening conditions for the assessment of sound programme material: multichannel sound, European Broadcast Union.

Einstein, M. A. (**1991**), *Sensory Science Theory and Applications in Foods*, in *IFT basic symposium series* Lawless and Klein (1991), chapter 11: Descriptive techniques and their hybridization, pp. 317–338.

Escofier, B. and Pagès, J. (**1988**). Analyses factorielles simples et multiples; Objectifs, méthodes et interprétation, forth edn, Paris: Dunod.

Escoufier, Y. (**1973**). "Le traitement des variables vectorielles," *Biometrics* **29**, 751–760.

Fastl, H. (**1997**). "The psychoacoustics of Sound-Quality Evaluation," *Acustica* **83**, 754–764.

Faye, P., Brémaud, D., Daubin, M. D., Courcoux, P., Giboreau, A., and Nicod, H. (**2004**). "Perceptive free sorting and verbalization tasks with naive subjects: An alternative to descriptive mappings," *Food Quality and Preference* **15**, 781–791.

Findlay, C. J., Castura, J. C., and Lesschaeve, I. (**2007**). "Feedback calibration: A training method for descriptive panels," *Food Quality and Preference* **18**, 321–328.

Fletcher, H. (**1934**). "An acoustic illusion telephonically achieved," *Bell Laboratories Record* **11**, 286–289.

Ford, N., Rumsey, F., and Nind, T. (**2002**). "Subjective evaluation of perceived spatial differences in car audio systems using a graphical assessment language," in *Proceedings of the 112nd Convention of the Audio Engineering Society*, Munich, Germany.

Frost, W. A. K. and Braine, R. L. (**1967**). "The application of the Repertory Grid technique to problems in market research," *Commentary* **9**, 161–175.

Gabrielsson, A. (**1979a**). "Dimension analyses of perceived quality of sound reproduction systems," *Scandinavian Journal of Psychology* **20**, 159–169.

Gabrielsson, A. (**1979b**). "Statistical treatment of data for listening tests on sound reproduction systems," Technical Report TA 92, Department of Technical Audiology, Karolinska Inst., Sweden.

Gabrielsson, A. and Sjögren, H. (**1979**). "Perceived sound quality of sound-reproduction systems," *Journal of the Acoustical Society of America* **65**, 1019–1033.

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

Gerzon, M. A. (**1973**). "Periphony: With height sound reproduction," *Journal of the Audio Engineering Society* **21**, 2–10.

Gerzon, M. A. (**1974**). "Surround sound psychoacoustics," *Wireless World* **80**, 483–486.

Giboreau, A., Dacremont, C., Egoroff, C., Guerrand, S., Urdapilleta, I., Candel, D., and Dubois, D. (**2007**). "Defining sensory descriptors: Towards writing guidelines based on terminology," *Food Quality and Preference* **18**, 265–274.

Gower, J. C. (**1975**). "Generalized procrustes analysis," *Psychometrika* **40**, 33–50.

Gower, J. C. and Dijksterhuis, G. B. (**2004**). Procrustes Problems, Oxford University Press.

Gower, J. C. and Hand, D. J. (**1996**). Biplots, CRC Press.

Greenhoff, K. and MacFie, H. J. H. (**1994**), *Measurement of Food Preferences*, London: Blackie Academic and Professional, chapter 6: Preference mapping in practice, pp. 137–166.

Gröhn, M., Lokki, T., and Takala, T. (**2002**). "An orientation experiment using auditory artificial horizon," in *Proceedings of the International Conference on Auditory Display*, ICAD, Kyoto, Japan, pp. 394–402.

Guastavino, C. and Katz, B. F. G. (**2004**). "Perceptual evaluation of multi-dimensional spatial audio reproduction," *Journal of the Acoustical Society of America* **116**, 1105–1115.

Guy, C., Piggott, J. R., and Marie, S. (**1989**). "Consumer profiling of scotch whisky," *Food Quality and Preference* **1**, 69–73.

Hammershøi, D. and Møller, H. (**2005**), *Binaural Technique – Basic Methods for Recording, Synthesis, and Reproduction*, Springer, chapter 9, pp. 223–254.

Hanafi, M., Qannari, E. M., Schlich, P., and Ledauphin, S. (**2004**). "Analysis of sensory profiling data by taking account of the performance of the assessors," in 7^{th} *Sensometrics Meeting*, Davis, CA, USA.

Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Hiipakka, J., and Lorho, G. (**2004**). "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society* **52**, 618–639.

Harshman, R. A. (**1972**). "PARAFAC2: Mathematical and technical notes," *UCLA working papers in phonetics* **22**, 30–44.

Hawkes, R. J. and Douglas, H. (**1971**). "Subjective acoustic experience in concert auditoria," *Acustica* **24**, 235–250.

Hegarty, P., Choisel, S., and Bech, S. (**2007**). "A listening test system for automotive audio – Part 3: Comparison of attribute ratings made in a vehicle with those made using an auralization system," in *Proceedings of the 123rd Convention of the Audio Engineering Society*, New-York, USA.

Heller, J. (**2000**). "Representation and assessment of individual semantic knowledge," *Methods of Psychological Research – Online* **5**, 1–37.

Hollowood, T. A. (**2004**). "Assessing panel performance," in *Nordic Workshop in Sensory Science, Turku, Finland*.

Huitson, A. (**1989**). "Problems with procrustes analysis," *Journal of Applied Statistics* **16**, 39–45.

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

Husson, F., Lê, S., and Pagès, J. (**2005**). "Confidence ellipse for the sensory profiles obtained by principal component analysis," *Food Quality and Preference* **16**, 245–250.

Husson, F., Lê, S., and Pagès, J. (**2007**). "Variability of the representation of the variables resulting from pca in the case of a conventional sensory profile," *Food Quality and Preference* **18**, 933–937.

Hynninen, J. and Zacharov, N. (**1999**). "Guinea Pig – A generic subjective test system for multichannel audio," in *Proceedings of the 106th Convention of the Audio Engineering Society*, Munich, Germany.

IEC 60268-13 (**1998**). Sound system equipment – Part 13: Listening tests on loudspeakers, International Electrotechnical Commission.

IEC 60268-7 (**1996**). Sound system equipment – Part 7: Headphones and earphones, International Electrotechnical Commission.

Isherwood, D., Lorho, G., Mattila, V.-V., and Zacharov, N. (**2003**). "Augmentation, application and verification of the generalized listener selection procedure," in *Proceedings of the 115th Convention of the Audio Engineering Society*, New York, USA.

Ishii, R., Chang, H. K., and O'Mahony, M. (**2007**). "A comparison of serial monadic and attribute-by-attribute protocols for simple descriptive analysis with untrained judges," *Food Quality and Preference* **18**, 440–449.

Ishii, R., Stampanoni, C., and O'Mahony, M. (**2008**). "A comparison of serial monadic and attribute-by-attribute descriptive analysis protocols for trained judges," *Food Quality and Preference* **19**, 277–285.

ISO 11035 (**1994**). Sensory analysis – Identification and selection of descriptors for establishing a sensory profile by a multidimensional approach, International Organization for Standards.

ISO 13299 (**2003**). Sensory analysis – Methodology – General guidance for establishing a sensory profile, International Organization for Standards.

ISO 4120 (**2004**). Sensory analysis – Methodology – Triangle test, International Organization for Standards.

ISO 5495 (**2005**). Sensory analysis – Methodology – Paired comparison test, International Organization for Standards.

ISO 5497-1 (**1987**). Precision of test methods – Part 1: Guide for the determination of repeatability and reproducibility for a standard test method by inter-laboratory tests, International Organization for Standards.

ISO 5725-1 (**1994**). Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions, International Organization for Standards.

ISO 8586-1 (**1993**). Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 1: Selected assessors, International Organization for Standards.

ISO 8586-2 (**1994**). Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 2: Experts, International Organization for Standards.

ISO 9000 (**2000**). Quality Management Systems – Fundamentals and vocabulary, International Organization for Standards.

ITU-R (**1994**). Recommendation BS.775-1, Multichannel stereophonic sound systems with and without accompanying picture, International Telecommunications Union Radiocommunication Assembly.

ITU-R BS.1116-1 (**1997**). Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, International Telecommunications Union Radiocommunication Assembly.

ITU-R BS.1284-1 (**2003**). General methods for the subjective assessment of sound quality, International Telecommunications Union Radiocommunication Assembly.

ITU-R BS.1534-1 (**2003**). Method for the subjective assessment of intermediate quality level of coding systems, International Telecommunications Union Radiocommunication Assembly.

ITU-T P.57 (**2002**). Telephone transmission quality. Objective measuring apparatus. Artificial ears, International Telecommunications Union, Telecommunications Standardization Sector.

ITU-T P.800 (**1996**). Methods for subjective determination of transmission quality, International Telecommunications Union, Telecommunications Standardization Sector.

ITU-T P.831 (**1998**). Subjective performance evaluation of network echo cancellers, International Telecommunications Union, Telecommunications Standardization Sector.

ITU-T P.832 (**2000**). Subjective performance evaluation of handsfree terminals, International Telecommunications Union, Telecommunications Standardization Sector.

Jack, F. and Piggott, J. (**1991**). "Free choice profiling in consumer research," *Food quality and preference* **3**, 129–134.

Jankowicz, D. and Thomas, L. (**1982**). "An algorithm for the cluster analysis of repertory grids in human resource development," *Personnel Review* **11**, 15–22.

Jeffress, L. A. and Taylor, R. W. (**1961**). "Lateralization vs localization," *Journal of the Acoustical Society of America* **33**, 482–483.

Johnson, D. N. (**2006**), Sensory testing and mechanical perceptions of quality – A novel application of quick individual vocabulary profiling, Master's thesis, Helsinki University of Technology, Helsinki, Finland.

Jolliffe, I. (**2002**). Principal component analysis, Springer verlag.

Kahle, G. (**1995**). *Validation d'un modéle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras*, PhD thesis, IRCAM, Paris, France.

Kelly, G. (**1955**). The psychology of personal constructs, Norton, New York.

Kelly, G. A. (**1963**). "Non-parametric factor analysis of personality theories," *Journal of Individual Psychology* **19**, 115–147.

Kiers, H. A. L., ten Berge, J. M. F., and Bro, R. (**1999**). "PARAFAC2 – Part 1. A direct fitting algorithm for the PARAFAC2 model," *Journal of Chemometrics* **13**, 275–294.

Kim, S. and Martens, W. L. (**2007**). "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.

King, B. M. and Arents, P. (**1991**). "A statistical test of consensus obtained from generalized procrustes analysis of sensory data," *Journal of Sensory Studies* **6**, 37–48.

Kirkeby, O. (**2002**). "A Balanced Stereo Widening Network for Headphones," in *Proceedings of the 22nd International Conference of the Audio Engineering Society*, Espoo, Finland, pp. 117–120.

Kistler, D. J. and Wightman, F. L. (**1992**). "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *Journal of the Acoustical Society of America* **91**, 1637–1647.

Kristof, W. and Wingersky, B. (**1971**). "Generalizations of the orthogonal procrustes rotation procedure to more than two matrices," in *Proceedings of the 79th Annual Convention of the American Psychological Association*, Vol. 6, pp. 89–90.

Kunert, J. and Qannari, E. L. M. (**1999**). "A simple alternative to generalized procrustes analysis: Application to sensory profiling data," *Journal of sensory studies* **14**, 197–208.

Kuttruff, K. H. (**1976**). Room Acoustics, Applied Science Publishers, London, UK.

Kylliäinen, M., Helimäki, H., Zacharov, N., and Cozens, J. (**2003**). "Compact high performance listening spaces," in *Proceedings of Euronoise*, Naples, Italy.

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

Larcher, V., Jot, J.-M., and Vandernoot, G. (**1998**). "Equalization methods in binaural technology," in *Proceedings of the 105th Convention of the Audio Engineering Society*, San Francisco, California, USA.

Lavandier, C. (**1989**). *Validation perceptive d'un modèle objectif de caractérisation de la qualité acoustique des salles*, PhD thesis, Université du Maine, Le Mans, France.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (**1994**). "The ACT (STATIS method)," *Computational statistics and data analysis* **18**, 97–120.

Lawless, H. T. (**1999**). "Descriptive analysis of complex odors: Reality, model or illusion?," *Food Quality and Preference* **10**, 325–332.

Lawless, H. T. and Heyman, H. (**1998**). Sensory Evaluation of Food: Principles and Practices, Chapman and Hall.

Lawless, H. T. and Klein, B. P. (**1991**). Sensory science theory and applications in foods, IFT basic symposium series, M. Dekker, New York, USA.

Lê, S., Josse, J., and Husson, F. (**2008**). "FactoMineR: An R package for multivariate analysis," *Journal of Statistical Software* **25**, 1–18.

Lea, P., Næs, T., and Rødbotten, M. (**1997**). Analysis of variance for sensory data, Johm Wiley and Sons, Chichester, UK.

Lea, P., Rødbotten, M., and Næs, T. (**1995**). "Measuring validity in sensory analysis," *Food Quality and Preference* **6**, 321–326.

Ledauphin, S., Hanafi, M., and Qannari, E. M. (**2006**). "Assessment of the agreement among the subjects in fixed vocabulary profiling," *Food Quality and Preference* **17**, 277–280.

Le Dien, S. and Pagès, J. (**2003a**). "Analyse factorielle multiple hiérarchique," *Revue de statistique appliquée* **51**, 47–73.

Le Dien, S. and Pagès, J. (**2003b**). "Hierarchical Multiple Factor Analysis: Application to the comparison of sensory profiles," *Food quality and preference* **14**, 397–403.

Letowski, T. (**1989**). "Sound quality assessment: Concepts and criteria," in *Proceedings of the 87th Convention of the Audio Engineering Society*, New York, USA.

L'Hermier Des Plantes, J. (**1976**). *Structuration des tableaux à trois indices de la statistique*, PhD thesis, Université de Montpellier II, Montpellier, France.

Lorho, G. (**2005a**). "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating," in *Proceedings of the 118th Convention of the Audio Engineering Society*, Barcelona, Spain.

Lorho, G. (**2005b**). "Individual Vocabulary Profiling of spatial enhancement systems for stereo headphone reproduction," in *Proceedings of the 119th Convention of the Audio Engineering Society*, New York, USA.

Lorho, G. (**2006**). "The effect of loudspeaker frequency bandwidth limitation and stereo base width on perceived quality," in *Proceedings of the 120th Convention of the Audio Engineering Society*, Paris, France.

Lorho, G. (**2007**). "Perceptual evaluation of mobile multimedia loudspeakers," in *Proceedings of the 122nd Convention of the Audio Engineering Society*, Vienna, Austria.

Lorho, G. (**2008**). "Analyzing four-way sensory data resulting from individual vocabulary profiling of audio: A comparison between HMFA and PARAFAC2," in *9th Sensometrics Meeting*, St. Catherines, Ontario, Canada.

Lorho, G. and Zacharov, N. (**2004**). "Subjective evaluation of virtual home theatre sound systems for loudspeakers and headphones," in *Proceedings of the 116th Convention of the Audio Engineering Society*, Berlin, Germany.

Lorho, G., Isherwood, D., Zacharov, N., and Huopaniemi, J. (**2002**). "Round robin

subjective evaluation of stereo enhancement systems for headphones," in *Proceedings of the 22nd International Conference of the Audio Engineering Society*, Espoo, Finland.

Lorho, G., Westad, F., and Bro, R. (**2006**). "Generalized correlation loadings – Extending correlation loadings to congruence and to multi-way models," *Chemometrics and Intelligent Laboratory Systems* **84**, 119–125.

Mackensen, P., Felderhoff, U., and Theile, G. (**1999**). "Binaural Room Scanning – A new Tool for Acoustic and Psychoacoustic Research," *Journal of the Acoustical Society of America* **105**, 1343–1344.

Majou, D., Touraille, C., and Hossenlopp, J. (**2001**). Sensory Evaluation – Guide of Good Practice, ACTIA, Imprimerie de l'Indre, Paris, France.

Mandel, J. (**1991**). "The validation of measurement through inter-laboratory studies," *Chemometrics and Intelligent Laboratory Systems* **11**, 109–119.

Mangan, P. A. P. (**1992**). "Performance assessment of sensory panelists," *Journal of Sensory Studies* **7**, 229–229.

Marchall, R. J. and Kirby, S. P. J. (**1988**). "Sensory measurement of food texture by free-choice profiling," *Journal of Sensory Studies* **3**, 63–80.

Martens, H. and Martens, M. (**2001**). Multivariate analysis of quality – An introduction, John Wiley.

Martens, W. L. (**1999**). "The impact of decorrelated low-frequency reproduction on auditory spatial imagery: Are two subwoofers better than one," in *Proceedings of the 16th International Conference of the Audio Engineering Society*, Rovaniemi, Finland, pp. 67–77.

Martens, W. L. and Zacharov, N. (**2003**). "Spatial distribution of reflections affects auditory quality and character of speech sounds located in a virtual acoustic environment," in *Proceedings of the 1st ISCA ITRW on Auditory Quality of Systems*, Akademie Mont-Cenis, Germany.

Marui, A. and Martens, W. L. (**2006**). "Spatial character and quality assessment of selected stereophonic image enhancements for headphone playback of popular music," in *Proceedings of the 120th Convention of the Audio Engineering Society*, Paris, France.

Mason, R., N Ford, F. R., and Bryun, B. (**2001**). "Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction," *Journal of the Audio Engineering Society* **49**, 366–384.

Mattila, V.-V. (**2001a**). "Descriptive analysis of speech quality in mobile communications: Descriptive language development and external preference mapping," in *Proceedings of the 111th Convention of the Audio Engineering Society*, New York, USA.

Mattila, V.-V. (**2001b**). *Perceptual analysis of speech quality in mobile communications*, PhD thesis, Tampere University of Technology, Tampere, Finland.

Mattila, V.-V. and Zacharov, N. (**2001**). "Generalized listener selection (GLS) procedure," in *Proceedings of the 110th Convention of the Audio Engineering Society*, Amsterdam, Holland.

Mazzucchelli, R. and Guinard, J. X. (**1999**). "Comparison of monadic and simultane-

ous sample presentation modes in a descriptive analysis of milk chocolate," *Journal of Sensory Studies* **14**, 235–248.

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. D., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological Research* **58**, 177–192.

McEwan, J. A. (**1996**), *Multivariate Analysis of Data in Sensory Science*, Elsevier, chapter 3: Preference mapping for product optimization, pp. 71–102.

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

McEwan, J. A., Hunter, E. A., van Gemert, L. J., and Lea, P. (**2002**). "Proficiency testing for sensory profile panels: measuring panel performance," *Food Quality and Preference* **13**, 181–190.

Meilgaard, M., Civille, G. V., and Carr, B. T. (**1991**). Sensory evaluation techniques, CRC Press.

Middlebrooks, J. C. (**1999**). "Individual differences in external-ear transfer functions reduced by scaling in frequency," *Journal of the Acoustical Society of America* **106**, 1480–1492.

Miller, J. and Carterette, E. (**1975**). "Perceptual space for musical structures," *Journal of the Acoustical Society of America* **58**, 711.

Monrozier, R. and Danzart, M. (**2001**). "A quality measurement for sensory profile analysis The contribution of extended cross-validation and resampling techniques," *Food Quality and Preference* **12**, 393–406.

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society* **45**, 224–239.

Morand, E. and Pagès, J. (**2006**). "Procrustes multiple factor analysis to analyse the overall perception of food products," *Food quality and preference* **17**, 36–42.

Munoz, A. M. and Civille, G. V. (**1998**). "Universal, product and attribute specific scaling and the development of common lexicons in descriptive analysis," *Journal of Sensory Studies* **13**, 57–75.

Murray, J. M., Delahunty, C. M., and Baxter, I. A. (**2001**). "Descriptive sensory analysis: past, present and future," *Food Research International* **34**, 461–471.

Myers, R. H. and Montgomery, D. C. (**2002**). Response Surface Methodology : Process and Product Optimization Using Designed Experiments, Series in Probability and Statistics, 2^{nd} edn, Wiley-Interscience.

Næs, T. (**1990**). "Handling individual differences between judges in sensory profiling," *Food Quality and Preference* **6**, 187–199.

Næs, T. and Langsrud, Ø. (**1998**). "Fixed or random assessors in sensory profiling?," *Food Quality and Preference* **9**, 145–152.

Næs, T. and Risvik, E. (**1996**). Multivariate analysis of data in sensory science, Vol. 16 of *Data handling in science and technology*, Elsevier, Amsterdam, The Netherlands.

Næs, T. and Solheim, E. (**1991**). "Detection and interpretation of variation within

and between assessors in sensory profiling," *Journal of Sensory Studies* **6**, 159–177.

Nakayama, T., Miura, T., Kosaka, O., Okamoto, M., and Shiga, T. (**1971**). "Subjective assessment of multichannel reproduction," *Journal of the Audio Engineering Society* **19**, 744–751.

Neher, T., Brookes, T., and Rumsey, F. J. (**2006**). "A hybrid technique for validating unidimensionality of perceived variation in a spatial auditory stimulus set," *Journal of the Audio Engineering Society* **54**, 259–275.

Neher, T., Rumsey, F. J., and Brookes, T. (**2002**). "Training of listeners for the evaluation of spatial sound reproduction," in *Proceedings of the 112th Convention of the Audio Engineering Society*, Munich, Germany.

Nilsen, A., Tomic, O., and Næs, T. (**2007**). "Use of PanelCheck to compare panel performance from inter collaborative test," in *7th Pangborn Sensory Science Symposium*, Minneapolis, USA.

Nunnally, J. C. and Bernstein, I. H. (**1994**). Psychometric theory, 3rd edn, McGraw-Hill, New York.

Nyman, G., Radun, J., Leisti, T., Oja, J., Ojanen, H., Olives, J.-L., Vuori, T., and Häkkinen, J. (**2006**). "What do users really perceive: Probing the subjective image quality," in *Proceedings of SPIE*, Vol. 6059, p. 605902.

Olson, J. C. (**1981**). "The importance of cognition processes and existing knowledge structures for understanding food acceptance," *Criteria of Food Acceptance. Forester Publishing; Zurich, Switzerland* pp. 69–81.

O'Mahony, M. (**1991**), *Sensory science theory and applications in foods*, in *IFT basic symposium series* Lawless and Klein (1991), chapter 8: Descriptive analysis and concept alignment.

O'Mahony, M. (**1995**). "Sensory measurement in food science: Fitting methods to goals," *Food technology* **49**, 72–82.

O'Mahony, M. and Rousseau, B. (**2003**). "Discrimination testing: A few ideas, old and new," *Food Quality and Preference* **14**, 157–164.

Osgood, C. E. (**1952**). "The nature and measurement of meaning," *Psychol Bull* **49**, 197–237.

Osgood, C. E. (**1967**). The Measurement of Meaning, University of Illinois Press.

Pagès, J. and Husson, F. (**2001**). "Inter-laboratory comparison of sensory profiles methodology and results," *Food quality and preference* **12**, 297–309.

Pagès, J. and Husson, F. (**2005**). "Multiple factor analysis with confidence ellipses: a methodology to study the relationships between sensory and instrumental data," *Journal of Chemometrics* **19**, 138–144.

Pavel, A. (**1978**). "High fidelity stereophonic reproduction system," U.S. patent no. 4,412,106.

Pedersen, T. H. and Fog, C. L. (**1998**). "Optimisation of perceived product quality," in *Euronoise 98*, Vol. II, Euronoise, pp. 633–638.

Pellegrini, R. (**2001**). *A virtual reference listening room as an application of auditory virtual environments*, PhD thesis, Institut für Kommunikationsakustik, Ruhr-Universität, Bochum, Germany.

Peltonen, T., Lokki, T., Gouatarbes, B., Merimaa, J., and Karjalainen, M. (**2001**). "A system for multi-channel and binaural room response measurements," in *Proceedings of the 110th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands.

Perrin, L., Symoneaux, R., Maître, I., Asselin, C., Jourjon, F., and Pagès, J. (**2008**). "Comparison of three sensory methods for use with the Napping® procedure: Case of ten wines from Loire valley," *Food Quality and Preference* **19**, 1–11.

Piggott, J. R. (**1991**), *Sensory science theory and applications in foods*, in *IFT basic symposium series* Lawless and Klein (1991), chapter 12: Selection of terms for descriptive analysis.

Piggott, J. R. and Watson, M. P. (**1992**). "Comparison of free-choice profiling and the repertory grid method in the flavor profiling of cider," *Journal of Sensory Studies* **7**, 133–145.

Piggott, J. R., Simpson, S. J., and Williams, S. A. R. (**1998**). "Sensory analysis," *International Journal of Food Science and Technology* **33**, 7–12.

Plenge, G. (**1974**). "On the differences between localization and lateralization," *Journal of the Acoustical Society of America* **56**, 944–951.

Popper, R. and Heymann, H. (**1996**), *Multivariate analysis of data in sensory science*, Vol. 16 of *Data handling in science and technology* Næs and Risvik (1996), chapter 6: Analyzing differences among products and panelist by multidimensional scaling, pp. 159–184.

Qannari, E. M., Wakeling, I., and MacFie, H. J. H. (**1995**). "A hierarchy of models for analysing sensory data," *Food quality and preference* **6**, 309–314.

Qannari, E. M., Wakeling, I., Courcoux, P., and MacFie, H. J. H. (**2000**). "Defining the underlying sensory dimensions," *Food Quality and Preference* **11**, 151–154.

Quesnel, R. (**1996**). "Timbral ear trainer: Adaptive, interactive training of listening skills for evaluation of timbre difference," in *Proceedings of the 98th Convention of the Audio Engineering Society*, Paris, France.

Quesnel, R. (**2002**). *A computer-assisted method for training and researching timbre memory and evaluation skills*, PhD thesis, McGill University, Montreal, Canada.

Quesnel, R. and Woszczyk, W. R. (**1994**). "A computer-aided system for timbral ear training," in *Proceedings of the 96th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands.

Rainey, B. A. (**1986**). "Importance of reference standards in training panelists," *Journal of Sensory Studies* **1**, 149–154.

Riederer, K. A. J. (**2005**). *HRTF analysis: Objective and subjective evaluation of measured head-related transfer functions*, PhD thesis, Helsinki University of Technology, Espoo, Finland.

Risvik, E., Colwill, J., McEwan, J., and Lyon, D. (**1992**). "Multivariate analysis of conventional profiling data: a comparison of a British and a Norwegian trained panel," *Journal of Sensory Studies* **7**, 97–118.

Risvik, E., McEwan, J., Colwill, J., Rogers, R., and Lyon, D. (**1994**). "Projective mapping: a tool for sensory analysis and consumer research," *Food quality and preference* **5**, 263–269.

Rivière, P., Saporta, G., Pagès, J., and Monrozier, R. (**2005**). "Kano's satisfaction model applied to external preference mapping: A new way to handle non linear relationships between hedonic evaluations and product characteristics," in *PLS05, 4^{th} International Symposium on PLS and related methods*, Vol. 158, Barcelona, Spain.

Robert, P. and Escoufier, Y. (**1976**). "A unifying tool for linear multivariate statistical methods: The RV-coefficient," *Journal of the Royal Statistical Society. Series C: Applied Statistics* **25**, 257–265.

Romano, R., Brockhoff, P. B., Hersleth, M., Tomic, O., and Næs, T. (**2008**). "Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis," *Food Quality and Preference* **19**, 197–209.

Rossi, F. (**2001**). "Assessing sensory panelist performance using repeatability and reproductibility measures," *Food Quality and Preference* **12**, 467–479.

Rumsey, F. (**1998**). "Subjective assessment of the spatial attributes of reproduced sound," in *Proceedings of the 15^{th} International Conference of the Audio Engineering Society*, Copenhagen, Denmark.

Rumsey, F. (**2001**). Spatial Audio, Focal Press, Oxford, UK.

Rumsey, F. (**2002**). "Spatial quality evaluation for reproduced sound: Terminology, meaning and a scene-based paradigm," *Journal of the Audio Engineering Society* **50**, 651–666.

Samoylenko, E., McAdams, S., and Nosulenko, V. (**1996**). "Systematic analysis of verbalisations produced in comparing musical timbres," *International Journal of Psychology* **31**, 255–278.

Sayers, B. M. and Cherry, E. C. (**1957**). "Mechanism of binaural fusion in the hearing of speech," *Journal of the Acoustical Society of America* **29**, 973–987.

Schlich, P. (**1994**). "GRAPES: A method and sas program for graphical representation of assessor performance," *Journal of Sensory Science* **9**, 157–169.

Schlich, P. (**1995**). "Preference mapping: Relating consumer preferences to sensory or instrumental measurements," in P. Etiévant and P. Schreier, eds, *Bioflavour'95*, INRA Editions, Paris, pp. 135–150.

Schlich, P. (**1996**), *Multivariate analysis of data in sensory science*, Vol. 16 of *Data handling in science and technology* Næs and Risvik (1996), chapter 9: Defining and validating assessor compromises about product distances and attribute correlations, pp. 259–306.

Schlich, P. (**1997**). "CAP: Une méthode et un outil de contrôle rapide et synthétique des performances des sujets en évaluation sensorielle descriptive," in *Proceedings of: 5^{emes} Journées Agroindustrie et Méthodes Statistiques, Versailles, France*, pp. 72–81.

Schlich, P. (**1998**). "What are the sensory differences among coffees? Multi-panel analysis of variance and FLASH analysis," *Food Quality and Preference* **9**, 103–106.

Schlich, P., Pineau, N., Brajon, D., and Cordelle, S. (**2006**). "Le projet Sensobase: Construire une base de profils sensoriels pour documenter les performances des panels d'analyse sensorielle," in *Proceedings of: 9^{emes} Journées Agroindustrie et Méthodes Statistiques, Montpellier, France.*

Schlich, P., Pineau, N., Brajon, D., and Quannari, E. M. (**2004**). "Multivariate control of panel performances," in *7th Sensometrics Meeting*, Davis, CA, USA.

Scriven, F. (**2005**). "Two type of sensory panel or are there more?," *Journal of Sensory Studies* **20**, 526–538.

Shaw, M. L. G. (**1980**). On becoming a personal scientist: Interactive computer elicitation of personal models of the world, Academic Press, London.

Sieffermann, J. M. (**2000**). "Le profil flash. un outil rapide et innovant d'évaluation sensorielle descriptive.," in *Agoral 2000 – XIIèmes rencontres "L'innovation: de l'idée au succès"*, Montpellier, France, pp. 335–340.

Silzle, A. (**2002**). "Selection and tuning of hrtfs," in *Proceedings of the 112nd Convention of the Audio Engineering Society*, Munich, Germany.

Sjöström, L. B. (**1954**), *Food Acceptance Testing Methodology*, Quartermaster Food and Container Institute, chapter : The descriptive analysis of flavor, pp. 25–61.

Slater, P. (**1977**). The measurement of interpersonal space by Grid Technique, Vol. 2, Wiley, London.

Smilde, A. K., Bro, R., and Geladi, P. (**2004**). Multi-way analysis with applications in the chemical sciences, John Wiley & Sons, Chichester, England.

Snow, W. B. (**1934**). "Auditory perspective," *Bell Laboratories Record* **12**, 194–198.

Soulodre, G. A., Grusec, T., Lavoie, M., and Thibault, L. (**1998**). "Subjective evaluation of state-of-the-art 2-channel audio codecs," *Journal of the Audio Engineering Society* **46**, 164–176.

Stampanoni, C. R. (**1993**). "The quantitative flavor profiling technique," *Perfumer & flavorist* **18**, 19–24.

Stampanoni, C. R. (**1994**). "The use of standardised flavour languages and quantitative flavour profiling technique for flavoured dairy products," *Journal of Sensory Studies* **9**, 383–400.

Steenkamp, J. B. and Trijp, H. V. (**1997**). "Attribute elicitation in marketing research: A comparison of three procedures," *Marketing Letters* **8**, 153–165.

Steinke, G. (**1996**). "Surround sound – The new phase: An overview," in *Proceedings of the 100th Convention of the Audio Engineering Society*, Copenhagen, Denmark.

Stevens, S. S. (**1961**). "Procedure for calculating loudness: Mark VII," *Journal of the Acoustical Society of America* **33**, 1577–1585.

Stone, H. and Sidel, J. L. (**1993**). Sensory evaluation practices, 2nd edn, Academic Press, San Diego, CA, USA.

Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, C. (**1974**). "Sensory evaluation by quantitative descriptive analysis," *Food Technology* **28**, 22–43.

Susini, J., McAdams, S., Winsberg, S., Perry, I., Vieillard, S., and Rodet, X. (**2004**). "Characterizing the sound quality of air-conditioning noise," *Applied Acoustics* **65**, 763–790.

Szczesniak, A. S. (**1963**). "Classification of textural characteristics," *Journal of Food Science* **28**, 385–389.

Szczesniak, A. S., Brandt, M. A., and Friedman, H. H. (**1963**). "Development of standard rating scales for mechanical parameters of texture and correlation between the

objective and the sensory methods of texture evaluation," *Journal of Food Science*.

Tarea, S., Cuvelier, G., and Sieffermann, J. M. (**2007**). "Sensory evaluation of the texture of 49 commercial apple and pear purees," *Journal of Food Quality* **30**, 1121–1131.

Tarea, S., Sieffermann, J. M., and Cuvelier, G. (**2003**). "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 *12th world Congress of Food Science and Technology*, Chicago, USA.

ten Berge, J. M. F. (**1977**). "Orthogonal Procrustes rotation for two or more matrices," *Psychometrika* **42**, 267–276.

Theile, G. (**1986**). "On the standardisation of the frequency response of high-quality studio headphones," *Journal of the Audio Engineering Society* **34**, 956–969.

Theile, G. (**2001**). "Multichannel natural music recording based on psychoacoustic principles," in *Proceedings of the 19th International Conference of the Audio Engineering Society*, Schloss Elmau, Germany, pp. 201–229.

Theile, G., Wittek, H., and Reisinger, M. (**2003**). "Potential wavefield synthesis applications in the multichannel stereophonic world," in *Proceedings of the 24th International Conference of the Audio Engineering Society*, Banff, Alberta, Canada.

Thomson, D. M. H. and McEwan, J. A. (**1988**). "An application of the repertory grid method to investigate consumer perceptions of foods," *Appetite* **10**, 181–193.

Thurstone, L. L. (**1927**). "A law of comparative judgment," *Psychological Review* **34**, 273–286.

Thybo, A. K. and Martens, M. (**2000**). "Analysis od sensory assessors in texture profiling of potatoes by multivariate modelling," *Food Quality and Preference* **11**, 283–288.

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

Toole, F. E. (**1970**). "In-head localization of acoustic images," *Journal of the Acoustical Society of America* **48**, 943–949.

Toole, F. E. (**1984**). "The acoustics and psychoacoustics of headphones," in *Proceedings of the 2nd International Conference of the Audio Engineering Society*, Anaheim, Califirnia, USA.

Toole, F. E. (**1985**). "Subjective measurements of loudspeaker sound quality and listener performance," *Journal of the Audio Engineering Society* **33**, 2–32.

Toole, F. E. (**1986**). "Loudspeaker measurements and their relationship to listener preferences: Part 2," *Journal of the Audio Engineering Society* **34**, 323–348.

Toole, F. E. (**1990**). "Subjective evaluation: Identifying and controlling the variables," in *Proceedings of the 8th International Conference of the Audio Engineering Society*, Washington, DC, USA.

Toole, F. E. and Sayers, B. M. A. (**1965**). "Lateralization judgments and the nature of binaural acoustic images," *Journal of the Acoustical Society of America* **37**, 319–324.

Tuomi, O. and Zacharov, N. (**2000**). "A real-time binaural loudness model," in *139th*

meeting of the Acoustical Society of America, Atlanta, USA.

Usher, J. and Woszczyk, W. (**2003**). "Design and testing of a graphical mapping tool for analyzing spatial audio scenes," in *Proceedings of the 24th International Conference of the Audio Engineering Society*, Banff, Alberta, Canada.

Wakeling, I. N., Raats, M. M., and MacFie, H. J. H. (**1992**). "A new significance test for consensus in generalized Procrustes analysis," *Journal of Sensory Studies* **7**, 91–96.

Wickelmaier, F. and Choisel, S. (**2005**). "Selecting participants for listening tests of multichannel reproduced sound," in *Proceedings of the 118th Convention of the Audio Engineering Society*, Barcelona, Spain.

Wickelmaier, F. and Ellermeier, W. (**2007**). "Deriving auditory features from triadic comparisons," *Perception and Psychophysics* **69**, 287–297.

Wightman, F. L. and Kistler, D. J. (**1989a**). "Headphone simulation of free-field listening. I: Stimulus synthesis," *Journal of the Acoustical Society of America* **85**, 858–867.

Wightman, F. L. and Kistler, D. J. (**1989b**). "Headphone simulation of free-field listening. II: Psychophysical validation," *Journal of the Acoustical Society of America* **85**, 868–878.

Wilkens, H. (**1975**). *Mehrdimensionale beschreibung subjektiver burteilungen der akustik von konzertsälen*, PhD thesis, TU Berlin.

Williams, A. A. and Arnold, G. M. (**1985**). "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**, 204–214.

Williams, A. A. and Langron, S. P. (**1984**). "The use of free-choice profiling for the examination of commercial ports," *Journal of the Science of Food and Agriculture* **35**, 558–568.

Wise, B., Gallagher, N., Bro, R., Shaver, J., Windig, W., and Koch, R. (**2006**), *PLS Toolbox 4.0 for use with MATLAB*, Eigenvector Research, Wenatchee, WA, USA.

Wold, S., Martens, H., and Wold, H. (**1983**). "The multivariate calibration method in chemistry solved by the PLS method," in B. K. A Ruhe, ed., *Matrix Pencils, Lecture Notes in Mathematics*, Springer-Verlag, Heidelberg, pp. 286–293.

Wolters, C. J. and Allchurch, E. M. (**1994**). "Effect of training procedure on the performance of descriptive panels," *Food Quality and Preference* pp. 203–214.

Zacharov, N. and Huopaniemi, J. (**1999**). "Results of a round robin subjective evaluation of virtual home theatre sound systems," in *Proceedings of the 107th Convention of the Audio Engineering Society*, New York City, NY, USA.

Zacharov, N. and Koivuniemi, K. (**2001a**). "Perceptual audio profiling and mapping of spatial sound displays," in *Proceedings of the International Conference on Auditory Display*, ICAD, Espoo, Finland.

Zacharov, N. and Koivuniemi, K. (**2001b**). "Unravelling the perception of spatial sound reproduction: Analysis & external preference mapping," in *Proceedings of the 111th Convention of the Audio Engineering Society*, New York City, NY, USA.

Zacharov, N. and Koivuniemi, K. (**2001c**). "Unravelling the perception of spatial

sound reproduction: Techniques and experimental design," in *Proceedings of the 19th International Conference of the Audio Engineering Society*, Schloss Elmau, Germany.

Zacharov, N. and Lorho, G. (**2006**). "What are the requirements of a listening panel for evaluating spatial audio quality," in *International Workshop on Spatial Audio and Sensory Evaluation Techniques*, University of Surrey, Guildford, UK.

Zielinski, S., Rumsey, F., and Bech, S. (**2008**). "On some biases encountered in modem audio quality listening tests: A review," *Journal of the Audio Engineering Society* **56**, 427–451.

Zwicker, E. and Fastl, H. (**1990**). Psychoacoustics: Facts and Models, Springer-Verlag, Heidelberg, Germany.

AALTO UNIVERSITY
SCHOOL OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF SIGNAL PROCESSING AND ACOUSTICS
REPORT SERIES