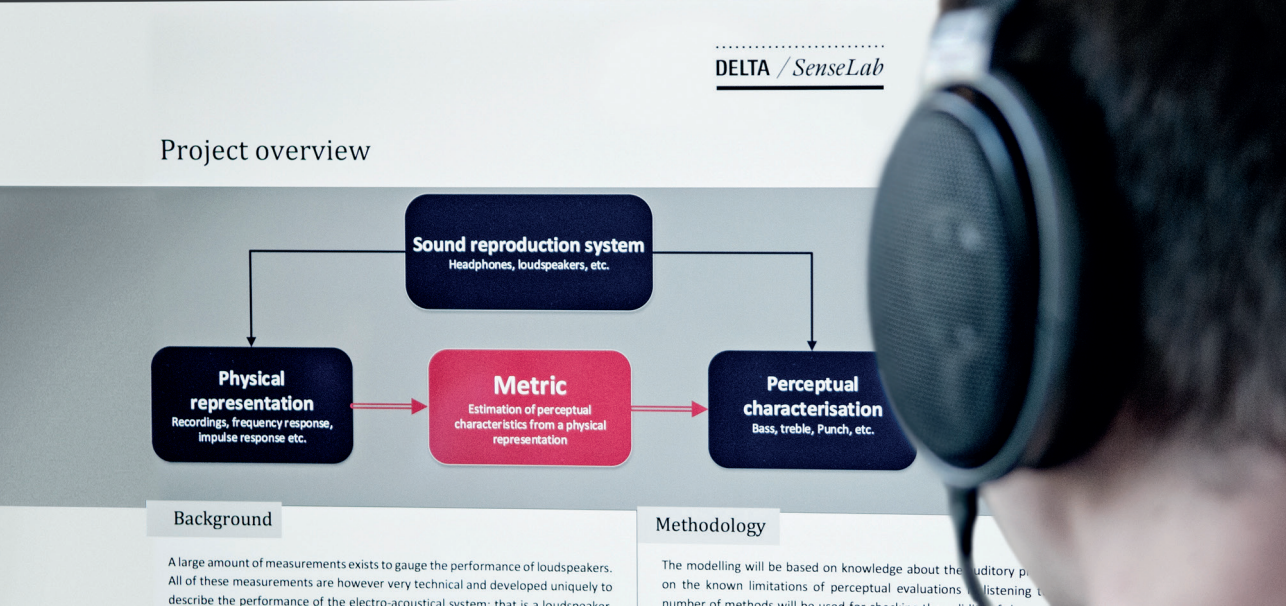
Prediction of perceptual audio reproduction characteristics

Volk, Christer Peter

DOI (link to publication from Publisher):
10.5278/vbn.phd.engsci.00164

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record



PREDICTION OF PERCEPTUAL AUDIO REPRODUCTION CHARACTERISTICS

BY
CHRISTER PETER VOLK

DISSERTATION SUBMITTED 2016

DELTA

Ministry of Higher Education and Science
—
Danish Agency for Science, Technology and Innovation

AALBORG UNIVERSITY
DENMARK

Prediction of perceptual audio reproduction characteristics

Ph.D. Dissertation

Christer Peter Volk

Dissertation submitted October 15th, 2016

Dissertation submitted:	October 15th, 2016
PhD supervisor:	Prof. Søren Bech Aalborg University
Assistant PhD supervisor:	Assoc. Prof. Flemming Christensen Aalborg University
Company Supervisor:	Torben Holm Pedersen DELTA SenseLab
PhD committee:	Associate Professor Christian Sejer Pedersen (chair.) Aalborg University, Denmark
	Associate Professor Tapio Lokki Aalto University School of Science, Finland
	Project Manager Lars Bramsløw Eriksholm Research Centre, Oticon, Denmark
PhD Series:	Faculty of Engineering and Science, Aalborg University

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-825-3

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2016

About the author

Christer P. Volk
Date of Birth: December 16^{th}*, 1984*
Nationality: Danish

Christer P. Volk received a B.Eng. degree in Electronics and Computer engineering from the Engineering College of Copenhagen in 2009. In his bachelor project he made an algorithm for improving speech intelligibility for hearing impaired students in classrooms.

He then went on to study acoustics, focusing on psychoacoustics, at the Technical University of Denmark, where he obtained an M.Eng. degree in Engineering Acoustics. In his master thesis, he investigated the relations between the hearing loss of hearing impaired subjects and their performance in perceptual evaluation of perceptual characteristics. The project included elements of psychoacoustics, sensory science, recording technique, as well as multivariate statistics. Simultaneously with the master programme he was an assessor in an expert listening panel, assessing a large variety of audio products and codecs.

In September 2013 he started on an industrial PhD at DELTA SenseLab in collaboration with Aalborg University, which this dissertation is a product of.

About the author

Abstract

Loudspeakers, headphones and other sound reproduction systems have traditionally been evaluated in two separate domains: The physical domain and the perceptual domain. The physical domain consists of technical measurements of e.g. frequency response, impulse response, directivity, sensitivity etc. and is traditionally measured in anechoic rooms. The perceptual domain consists of perceptual evaluation of stimuli in listening tests with highly controlled experimental variables, and is traditionally conducted in listening rooms (for loudspeakers) with a low and standardized reverberation time or in listening booths (for headphones); both with low background noise. The limited direct connection between these two domains of evaluation, have made it difficult to understand the perceptual consequence of the physical measurement results; Both in terms of overall preference and in terms of sound reproduction characteristics.

In this project the feasibility of modelling perceptual characteristics of headphones and loudspeakers was investigated, by establishing a number of prediction models and evaluating their prediction capabilities. An initial part of this investigation concerned strategies for obtaining a good data basis, i.e. an objective perceptual characterisation of a set of sound reproduction systems and physical data of relevance for the auditory perception. This investigation led to a strategy of obtaining perceptual and physical measurements in the listening position, used in the conducted listening tests, in an effort to optimize the direct connection between these two domains. Accordingly, the proposed methodology consisted of making physical measurements in the listening position using a Brüel & Kjær head-and-torso simulator with microphones placed at the same position as the human ear drums.

The human auditory processing is highly non-linear in terms sound pressure level as well as spectral, temporal and binaural properties of the sound reaching the ears. Because of that, the features of the unprocessed recorded stimuli were still not directly connected to human perception. Consequently, these recordings were processed using auditory models in order to obtain a perceptually relevant physical representation of the sound reproduction in the listening position.

A number of prediction models were proposed on the basis of the measurements from these two domains and each model was trained for prediction of one perceptual characteristic. Among these were both dominating characteristics differentiating headphones and commonly found sound-reproduction characteristics selected from a set of well-defined sensory descriptors known as a Sound wheel. The modelled characteristics comprised exclusively of spectral properties of the reproduction, but other types of characteristics were included in the investigations as well.

One generic prediction model was proposed, which was trained to accurately predict a number of sensory descriptors, such as Bass and Treble strength, while another type of model was designed specifically for prediction of the sensory descriptor Dark-Bright (a spectral-balance property). In total the modelling efforts led to 12 prediction metrics, which correlated well with results of the perceptual evaluations ($r = [0.70 - 0.99]$).

Resumé

Højttalere, hovedtelefoner og andet udstyr til gengivelse af lyd, er traditionelt set blevet evalueret i to separate domæner: Det fysiske domæne og det perceptuelle domæne. Det fysiske domæne består af tekniske målinger af f.eks. frekvensrespons, impulsrespons, retningsvirkning, sensitivitet, m.m. og måles traditionelt i lyddøde rum. Det perceptuelle domæne består af evalueringer af lydopfattelsen af stimuli repræsenteret i lyttetest med nøje kontrollerede eksperimentelle faktorer og bliver som oftest afholdt i lytterum (ved evaluering af/over højttalere) med en lav og standardiseret efterklangstid eller i lyttebokse (ved evaluering af/over hovedtelefoner); begge med lav baggrundsstøj. Den begrænset direkte sammenhæng mellem disse to domæner, har gjort det besværligt at forstå de perceptuelle konsekvenser af de fysiske måleresultater; Både i forhold til den overordnede præference, samt i forhold til karakteristika af lydgengivelsen.

I dette projekt blev der undersøgt, i hvor høj grad det var muligt at forudsige perceptuelle karakteristika af hovedtelefoner og højttalere. Dette blev gjort ved at udvikle en række modeller til forudsigelse af disses karakteristika, som efterfølgende blev evalueret i forhold til deres evne til korrekt at forudsige intensiteten af lytteres bedømmelser. En indledende del af denne undersøgelse omhandlede strategier for indsamling af et godt datagrundlag bestående af en objektiv perceptuel karakterisering af udstyr til lydgengivelse, samt fysisk måledata med relevans for lydopfattelsen. Denne undersøgelse førte til en strategi om at foretage fysiske målinger af stimuli i lyttetestenes lytteposition, således at den direkte sammenhæng mellem disse to domæner blev optimeret. Den forslåede metodik til dette formål blev derfor, at foretage disse fysiske målinger i lyttepositionen ved brug af en Brüel & Kjær mannequin med mikrofoner placeret i samme positioner, som menneskers trommehinder.

Den menneskelige hørelse er stærkt ulineær i forhold til opfattelse af lydstyrke, samt spektrale, temporale og binaurale egenskaber af den lyd, som når ørene. Derfor er der stadig ikke direkte sammenhæng mellem optagelsernes egenskaber og de perceptuelle karakteristika. Som konsekvens heraf, blev optagelserne efterfølgende processerede via en auditiv model af

hørelsen, således at der blev opnået en repræsentation af udstyr til lydgengivelse i det fysiske domæne, som var perceptuelt relevant for lytternes opfattelse af stimuli.

Et antal modeller blev udviklet på basis af målinger fra de to domæner og hver af disse modeller blev udviklet specifikt til forudsigelse af én opfattede karakteristika. Blandt disse var både dominerende karakteristika til differentiering af hovedtelefoner, samt typiske karakteristika for udstyr til lydgengivelse udvalgt fra et såkaldt lydhjul af veldefinerede sensoriske termer. De modellerede sensoriske termer bestod udelukkede af spektrale termer, men andre typer karakteristika var også inkluderet i projektets undersøgelser.

Én generel model blev udviklet, optimeret til præcisionsforudsigelser af flere karakteristika, så som Bass- og Diskantstyrke. Derudover blev en anden type model udviklet specifikt til forudsigelsen af bedømmelsen af det sensoriske term Mørk-Lys (karakteristika vedrørende spektral-balance). Totalt set, ledte projektets undersøgelser til 12 såkaldte metrikker, som alle korrelerede godt med bedømmelserne fra de perceptuelle evalueringer ($r = [0.70 - 0.99]$).

Contents

Thesis Details

Thesis Title:	Prediction of perceptual audio reproduction characteristics
Ph.D. Student:	Christer Peter Volk
Supervisors:	Prof. Søren Bech, Aalborg University Assoc. Prof. Flemming Christensen, Aalborg University Senior Technology Specialist Torben Holm Pedersen, DELTA SenseLab

The main body of this thesis consist of the following conference papers and journal articles.

[A] **C. P. Volk**, S. Bech, T. H. Pedersen, and F. Christensen, "Five aspects of maximizing objectivity from perceptual evaluations of loudspeakers: A literature study," *in Proc. of the Audio Engineering Society Convention 138*. Warsaw, Poland: Audio Engineering Society, May 2015, pp. 1–12, Convention paper 9230.

[B] **C. P. Volk**, M. Lavandier, S. Bech, and F. Christensen, "Identifying the dominating perceptual differences in headphone reproduction," *Submitted, J. Acoust. Soc. Am.*, Feb. 2016.

[C] **C. P. Volk**, T. H. Pedersen, S. Bech, and F. Christensen, "Modelling perceptual characteristics of prototype headphones," *in Proc. of the AES International Conference on Headphone Technology*, Aalborg, Denmark: Audio Engineering Society, Aug. 2016, pp. 1–9, Paper no. 5-2.

[D] **C. P. Volk**, S: Bech, T. H. Pedersen, and F. Christensen, "Modelling perceptual characteristics of loudspeaker reproduction in a stereo setup," *Submitted, Journal of the Audio Engineering Society*, Sep. 2016.

In addition to the main papers, one additional research study (see Appendix) was also made as well as a minor contribution to a conference paper by a master student, Sune Olsen, at DELTA (not inserted in this thesis).

[1] **C. P. Volk**, T. H. Pedersen, S. Bech, and F. Christensen, "Characterisation of acoustical environments: Physics and perception," *Ph.D. thesis appendix*, pp. 135–161, 2015.

[2] S. L. Olsen, F. Agerkvist, E. MacDonald, T. Stegenborg-Andersen, and **C. P. Volk**, "Modelling the perceptual components of loudspeaker distortion," *in Proc. of the Audio Engineering Society Convention 140*, Paris, France: Audio Engineering Society, Jun. 2016, pp. 1–9, Convention Paper 9549.

This thesis has been submitted for assessment in partial fulfilment of the PhD degree. The thesis is based on the submitted or published scientific papers which are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Preface

"Top speed, 0-60, they are just numbers. They are meaningless in themselves. What matters is whether they add up into a sensation, and this delivers a tremendous sensation! [Tires screeches]"

- James Daniel May, Top Gear (TV show), Season 21 (2014), Episode 4.

This thesis is submitted to the Doctoral School of Engineering and Science at Aalborg University in partial fulfilment of the requirements for the degree of Doctor of Philosophy. It falls within the framework of an industrial PhD project, which was a collaboration between Aalborg University and DELTA SenseLab, where the student was employed throughout the duration of the project.

The work was funded by DELTA and the Danish Agency for Science, Technology and Innovation (Case number: 1355-00061). All work was carried out in the period from September 1st, 2013 to the date of the thesis submission at DELTA, the Department of Electronic Systems at Aalborg University, as well as Ecole Nationale Des Travaux Publics De L'Etat (ENTPE) part of the University of Lyon in France (Spring of 2015).

Industrial involvement

This PhD project ran in parallel with a government-funded research project carried out by DELTA SenseLab on perceptual evaluation of reproduced sound. During this period colleagues in SenseLab had a number of activities, which directly influenced this project:

- Trained a panel of expert listeners specifically in evaluation of sound reproduction characterisation,

- conducted a number of sub-projects demonstrating the possibilities of developed evaluation techniques,
- developed a Sound wheel of sensory descriptors for full characteristics of sound reproduction systems, and
- designed two loudspeaker spinners for loudspeaker evaluations.

The panel of trained assessors was used for all tests described in this thesis, with the exception of the experiment presented in Paper B conducted in Lyon, France. These tests were all designed with the dual purpose of advancing the investigations of the government project as well as providing data for this PhD project. Consequently, the tests were not designed strictly for testing of hypotheses related to the PhD work. The experimental designs were made in a collaboration between the PhD student and SenseLab colleagues to best fit the dual purposes.

The sensory descriptors modelled in this thesis all originates from a Sound wheel, which the PhD student was peripherally involved in the development of, but which was headed by the company supervisor Torben H. Pedersen. He also designed the loudspeaker spinners used in the listening test presented in Paper D.

Besides the internal involvement from SenseLab colleagues, this project benefited from support from Danish companies lending equipment for testing purposes. This included loudspeakers from DALI, headphones from Aiaiai, as well and compact loudspeakers from the Nordic retail chain Hi-Fi Klubben.

Project focus

The scope of this PhD project was on modelling of a selection of sound reproduction systems and scenarios. This included:

- evaluation of headphones, as well as mono- and stereo setups of loudspeakers in the horizontal plane.
- evaluation in small rooms (mimicking living rooms).
- evaluation of a physical products' influence on the reference stimuli, but not artefacts from codecs, wireless streaming or the like.
- evaluation of full-range loudspeakers covering the majority of the audible frequency range, i.e. not subwoofers.

Acknowledgement

First of all, I would like to thank Søren V. Legarth, Head of Department at DELTA SenseLab, for believing in me, when deciding to invest in an industrial PhD project on perceptual modelling. In general, the SenseLab team have been immensely helpful and supportive during these last three years. I owe special thanks to Tore Stegenborg-Andersen for the hard work he has put into all the listening tests conducted at DELTA.

Furthermore, I would like to thank my supervisors at Aalborg University, Søren Bech and Flemming Christensen, for great and continuous guidance in all aspects of the work and life as an upcoming researcher. Their input have greatly influenced the direction and quality of both the scientific endeavours and the communication of it. The same goes for my company supervisor Torben H. Pedersen, who supported my work both with guidance as well as novel tools to improve on the quality of the thesis work, such as the Sound wheel and two ingenious loudspeaker spinners. His ideas has led to truly novel concepts in this project.

Additionally, I am very thankful to Mathieu Lavandier, ENTPE, Lyon, France, for an inspiring collaboration. His enthusiasm pushed me to do more and his attention to details, made me a better researcher and communicator. The work environment at ENTPE was made great by the wonderful and supportive people at the laboratory.

Finally, I would like to thank my family, my friends, and my colleagues at Aalborg University for supporting me during this period. Especially, my mother for her never-ending faith in me, and my good friend Jakob for indulging me in taking a time-out from the project to climb Africa's tallest mountain Kilimanjaro. It truly taught me to take on each challenge one step at a time.

Christer P. Volk
DELTA SenseLab & Aalborg University, October 15, 2016

Acronyms & glossary

Acronyms

ANOVA analysis of variance

D/R direct-to-reverberant

DoE design of experiments

ERP ear reference point

FEM finite element method

HATS head-and-torso simulator

ITU International Telecommunication Union

JND just-noticeable difference

LSD least significant difference

LOO leave-one-out

MDS multi-dimensional scaling

RMSE root-mean-square error

PCA principal component analysis

SEAP specialized expert assessor panel

SNR signal-to-noise ratio

STEP spectro-temporal excitation pattern

Glossary

A number of words and terms in this thesis were used with a specific definition in mind. These are defined in the list below:

Auralization Generally defined as the process of (re)creating a sound event, real or virtual, in another setting, e.g. recreating a choir singing in the reverberant environment of a church in a smaller listening room. In this thesis, it refers specifically to recreation of headphone sound reproduction, from measurements of their frequency response or complex transfer function. This process is sometimes referred to as virtualization (recreation of a physical product's sound reproduction).

Metric An equation for determining the perceptual intensity of a sensory descriptor on the basis of recordings or physical measurements.

Perceptual characteristic A perceptual attribute (defined below) of noticeable prominence. Definition adapted from [65].

Perceptual attribute An unique/independent property that can be perceived (perceptual, affective or connotative); It may or may not be prominent. Definition from [66].

Prediction model Describes a model in development, which in it's final form is referred to as a metric. A generic prediction model may be trained for e.g. two product groups, compact loudspeakers and headphones, leading to two metrics measuring the perceptual intensity of a sensory descriptor in their respective context.

Sensory descriptor A word or phrase that describes, identifies, or labels a perceptual characteristic of a system, e.g. the sound reproduction of a loudspeaker. Definition adapted from [66]. Note that a sensory descriptor describes one or more perceptual attributes, although in the ideal case a sensory descriptor would describe exactly one perceptual attribute.

Stimuli Generally, "stimuli maybe anything that evokes a response from an assessor when presented with the stimuli" [65]. In the context of this thesis, stimuli is specifically the reproduced sound reaching a listener's ears (ear reference point (ERP)) and which comprises the basis of the listener's perceptual evaluation.

System A system refers to any type of sound reproduction equipment, e.g. a loudspeaker or a set of headphones. Often used in the context of describing listening tests, where systems are the equipment being evaluated. Note that 'system' have a special meaning in the Appendix report (defined in Section 1, p. 135).

Part I

Introduction

Prediction of perceptual audio reproduction characteristics

1 Introduction

In December 1915 the modern moving-coil loudspeaker was introduced to the world by Peter L. Jensen (Dane) and Edwin S. Pridham (American) [9]. After its 1915 presentation in San Francisco to an amazed audience its popularity increased rapidly and loudspeakers were manufactured and sold in large parts of the world within the next decade. Fast forward to today, no modern house hold is without numerous loudspeakers in stereos, radios, computers, flat screen TVs, cars, mobile phones etc. While the technical concept of most modern loudspeakers has remained the same as 100 years ago, the sound quality of both the recording and the reproduction has increased significantly; Driven by numerous improvements in technology, materials, and understanding of the interaction between electrical and mechanical parts and the corresponding acoustical output.

Sound reproduction can fundamentally be divided into three components as depicted in Fig. 1. An electro-mechanical domain representing e.g. a loudspeaker, which moves a diaphragm and produces an acoustical output. This output is transmitted to a listener who will experience an auditory sensation. At present time, the electro-mechanical domain and the link to the acoustical

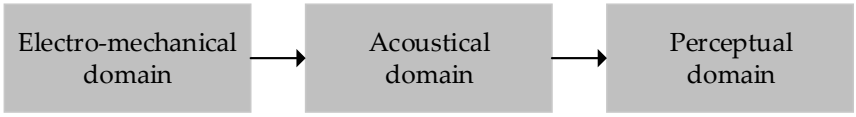


Fig. 1: The three fundamental domains of sound reproduction.

domain is well understood and we are able to calculate the acoustical output

of an electro-mechanical design with high precision using e.g. lumped element models [47, 74, 75] and finite element method (FEM) (e.g. [36]). How the acoustical output is perceived is, however, not yet as well understood. As a result, sound reproduction systems are currently described and gauged directly by physical properties: Frequency response, dimensions and volume of the cabinet, number and size of drivers, impedance, total harmonic distortion, sensitivity etc. An issue with these specifications is that they have very limited direct connection with the auditory sensation.

Traditionally, the sound character of loudspeakers have been (fine-)tuned using "golden ears" or tonmeisters. While this approach may have its merits, perceptual evaluations of a larger group of listeners, would be more representative of the average consumer perception than that of one tonmeister or a small group of experts. Obtaining information about perceptual characteristics of a more representative nature is, however, a time-consuming task involving complex listening tests, which may not be ideal in the early phases of development or for special use cases. One example could be the increasing amount of sound systems with adaptive acoustical outputs, where perceptual evaluations involving human listeners, might not be appropriate, if fast perceptual characterisation is needed.

Prediction of perceptual characteristics of sound reproduction systems would improve the possibility of setting perceptually-driven design goals, making e.g. a tonmeister able to specify perceptual characteristics as perceived by the average listener (or a target consumer segment), without relying on solely on their own senses. Having specifications based on predicted perceptual characteristics readily available would also allow decision-making in the engineering process on the basis of how a choice would affect perception. For example how the influence of a choice of speaker cone shape would affect perception of 'brilliance' and 'envelopment' rather that only knowing the affect on e.g. off-axis frequency response. This could be more relatable and especially useful in terms of making compromises. It is the potential of these predictions of auditory perception that led to the investigations described in this thesis.

The purpose of this PhD project was to establish and investigate the potential of mathematical models for prediction of perceived characteristics of the sound reproduction of headphones and loudspeakers. Specifically, whether a prediction model could be established that would be able to predict the intensity of a characteristic, i.e. the average rating given by a panel of experienced and trained listeners in listening tests; With the perceptual characteristics being represented by well-defined sensory descriptors (see definition in the Glossary, p. xx). Moreover, the purpose was to establish the requirements and limitations of current perceptual evaluation methods, with respect to obtaining an optimum data basis for establishing prediction

models. In this regard, it was of interest to establish how to predict listeners' ratings that are known to be relative in nature rather than absolute (see e.g. [1, 69]).

An important point with regards to the nature of the prediction models, was that perceptual characteristics sought modelled, was not the characteristics of listening test stimuli[1], but the differences in reproduction introduced by each set of headphones or each loudspeaker, i.e. the perceptual characteristics differentiating a listener's auditory sensation of a set of stimuli.

A key part of this project, was that the majority of perceptual evaluations (listening tests) were conducted with trained expert assessors, which were experienced in making perceptual evaluations and furthermore trained specifically in evaluation of perceptual characteristics of reproduced sound. Since the accuracy of the prediction models relied on the quality of the perceptual evaluations, emphasis was put on data quality analysis to ensure sufficient quality and suitability of data. Although it is possible to obtain quality data in perceptual evaluations with naïve listeners (consumers), the amount of subjects required is usually much higher. As an example of this, consider a sensory study of perfume characterisation [82], where a panel of 12 expert perfume assessors provided perceptual ratings with the same level of precision (uncertainty) as a panel of 103 naïve assessors (consumers), despite the large difference in panel sizes.

In the next section, previous modelling efforts within the domain of sound reproduction systems are summarised and important perspectives and challenges are highlighted. This is followed by a chapter on auditory processing, which describes why and how auditory perception was incorporated into the modelling methodology framework of this project as a key element. Chapter 3 provides a brief introduction into the two main types of perceptual measurements and their relation to objectivity as well as a description of a Sound wheel [66], which contains all of the sensory descriptors evaluated in this project. In Chapter 4, the issue of predicting ratings, which are dependent on the experimental setup (i.e. relative ratings) is discussed in detail. Additionally, an overview is given on the structure of the proposed predictions models, followed by a short section that introduces methods of gauging the performance of these. Finally, in Chapter 5, the contributions of this project are presented, discussed and summarised.

1.1 Previous modelling efforts

The pursuit for methods of measuring the dominating perceptual characteristics of audio reproduction span more than 40 years, pioneered by Gabrielsson in the 1970's and Toole in the 1980's. Since then many have followed in

[1]Stimuli is defined here, as the reproduced sound reaching a listener's ears. See full definition in the glossary, page xx.

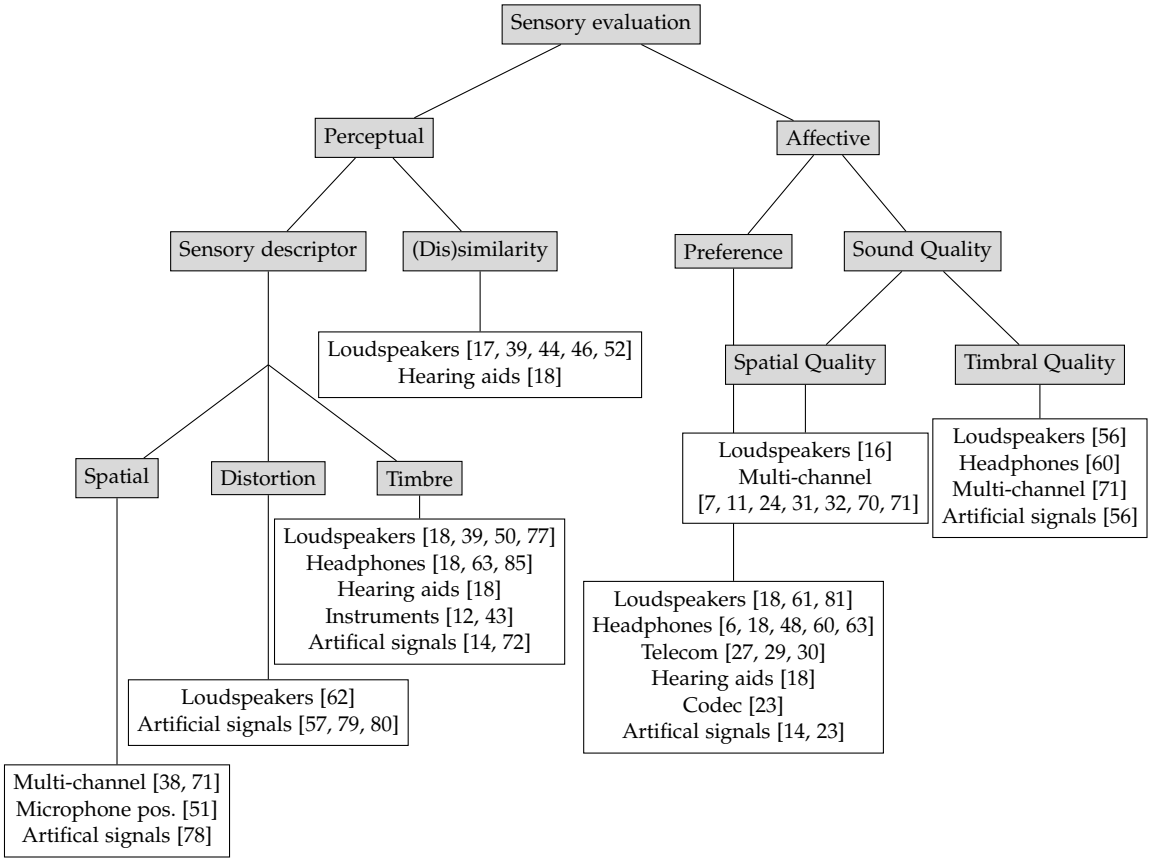


Fig. 2: Literature overview. The tree structure contains literature dealing with modelling of sensory ratings or evalutions of sensory characteristics relevant for modelling. The grey boxes contain classification of the studies, grouped into types of measurements (Perceptual and Affective) and subtypes of measurements. The white boxes contain the type of systems modelled and references to studies. Note that a source can appear in more than one box.

their path and the rate of publications suggest that the area have yet to peak. Among these efforts have been a number of attempts of perceptual modelling. In Fig. 2 an overview of these and related efforts are shown, categorised by domain, starting with the division into the perceptual and the affective domain. The overview emphasises the two main research areas: 1) subjective preference modelling and 2) prediction of timbral characteristics. Note that the Preference box includes studies with both Mean opinion score (MOS), Basic Audio Quality (BAQ), and Preference, i.e. all subjective terms without characterisation of any specific area of audio reproduction. Additionally, the research area of spatial evaluation and modelling of multi-channel and 3D sound for virtual- and augmented reality have been very active in recent years, and may not be fully represented by the literature referenced here.

The following subsections will briefly discuss a number of topics in the literature identified as important for understanding the paths chosen in the this PhD project; Many of which are discussed in more detail in the enclosed collection of papers. First, a discussion of why sensory characterisation is of value as a supplement to hedonic evaluations of preferences, and therefore of value to make predictions of. Secondly, the dominance of timbral characteristics for differentiating sound reproduction systems is discussed. A timbral dominance was found to throughout the studies of this project and in this second subsection, this trend is put into perspective by summarising the dominance of timbrel aspects in comparable studies. Thirdly, an assumption, made in the majority of published modelling efforts are discussed: that one prediction model is sufficient for accurate predictions, regardless of the type of sound reproduction system. While perception of e.g. bass might be the same regardless of the type of sound reproduction system, the data on which the models are trained, listener ratings, could be affected by the type of systems evaluated. This issue is therefore adressed in the third subsection. Fourthly, the efforts, taken in literature, to obtain physical measurements of relevance for the perceptual characteristics being predicted are discussed, which is a key element for obtaining predictions with high performance and robustness [23]. The fifth topic concerns the influence of the number of systems, stimuli, and listeners in listening tests on the generalisability of perceptual evaluations and is of value for optimising the validity of the data basis on which the prediction models are trained. Finally, the last topic discusses what listeners use as a reference, when making perceptual evaluations, e.g. how a listener determines whether a presented stimuli has a *'neutral'* amount of bass or *'a lot'* of bass? This philosophical question have had influence on both previously chosen modelling approaches as well as the efforts in the current project and constitutes an important aspect of estimating the auditory processing of listeners on a higher cognitive level.

Value of sensory evaluations beyond preference ratings

While studies of preference offers information of direct interest for manufacturers of audio reproduction systems, supplementing with perceptual measurements offer an insight into the underlying reasons behind listener preference. Evaluation of perceptual characteristics is often done, by rating a number of sensory descriptors (words) [1], defined either by each listener individually or in consensus among a number of listeners. These descriptors are chosen to represent perceptual characteristics of importance specifically for evaluation of a chosen set of sound reproduction systems under evaluation. They thereby provide an intuitive mean for communicating the perceptual characteristics and differences between sound reproduction systems. In combination with subjective evaluations of preference, perceptual characterisation adds the possibility of optimizing products beyond the current state-of-the-art, e.g. using the Ideal Profile method [83], which allow listeners to score the intensity of stimuli for a set of sensory descriptors, while also stating their preferred intensity for each of the descriptors.

Dominance of timbral characteristics

The emphasis on timbral studies within perceptual measurements are well founded. The most dominating perceptual characteristics in audio reproduction have been found to concern timbre and in particular spectral differences (See e.g. [46]. While studies of sensory descriptors needed for discrimination between audio reproduction systems have lead to a large number of sensory descriptors, the importance of each is rarely discussed. One example is however the headphone study by Olive et. al [60], which showed 'Good Spectral Balance' to have the highest correlation with preference ($r = 0.92$). The importance of timbre is discussed in detail in Paper A and investigated for headphones in Paper B.

"One size fits all" modelling assumption

The referenced studies of Fig. 2 all have in common that they rely on an assumption of *"one size fits all"*; Meaning that one prediction model is considered sufficient for making accurate prediction for all listeners, all technologies, price ranges, etc. It is, however, well-established in this and other domains that listeners do not have the same preference, but that preference tend to be divided into clusters (see e.g. [42, 73]. Some may for instance prefer extra bass and others extra treble. System characteristics may also be evaluated differently for subgroups of sound reproduction systems. For example in terms of bass strength, where the frequency range consider by listeners when evaluating small portable loudspeakers may differ from the frequency range considered when evaluating larger floor-standing loudspeakers. As

a result, a model intended for prediction of characteristics or preference of any type of sound reproduction systems, under the *"one size fits all"* assumption, may suffer in terms of prediction capabilities. In view of this pitfall, the two loudspeaker studies of this PhD project (Appendix in Part III and Paper D), were designed to include only loudspeakers within *one* category of loudspeakers per study. On the contrary, the headphone study in Paper B focused on modelling the dominating differences between headphones in general and thus was not limited to one specific type of headphones. A cluster analysis of the perceptual evaluations did not show any clear clustering between headphones. This indicated that the performance of the metrics would not be negatively affected by being trained on data from all of them.

In general, results (of a prediction model) are of course only valid for the systems tested in a given study, but specifically regarding prediction models, they are only of value for prediction of systems which were not. This makes validation of prediction models of high importances. The referenced studies in Fig. 2 are generally lacking in this area and many does not include validation at all (discussed further in Section 4.4).

Data basis: Physical measurements with perceptual relevance

A common trend in studies over last 40 years is a shift away from using physical measurements directly in prediction models. In the early studies, frequency responses were for instance widely used for prediction of preference or timbral characteristics. This was still the case in the late 1980's and early 1990's [16, 81]. In 1984 Staffeldt [77] showed the importance of taking the transfer function of the human head into account when evaluating timbral aspects, instead of using frequnecy responses directly. These were, however, still based on traditional measurements made in anechoic chambers. In 1991 Gabrielsson found that measurements made in a listening room correlated better with listener ratings of sound quality of loudspeakers than measurements in an anechoic chamber (or a reverberation room). Later on, in 2004 [61] included the influence of the listening room in his predictive model of loudspeaker preference, but the influence of the head, torso and outer ear were still not taken into account and neither were the many non-linearities of human auditory perception. This was, however, done by Klippel in his PhD project (see the 1990 thesis summary paper [39]), who calculated the frequency spectrum of the stimuli directly in the position where listeners were position during the perceptual evaluations. Furthermore, these spectra were processed through a stationary loudness model and used as input for prediction of a number of sensory descriptors for characterisation of loudspeakers. The use of auditory models have since been increasing in popularity and have been used for e.g. prediction of sensory characteristics within the spatial domain [78] and in understanding the dominating perceptual differences

between loudspeakers [45, 46].

In conclusion, evaluation of loudspeakers have moved away from strict laboratory environments and towards evaluation resembling real-world conditions. Additionally, perceptional measurements have become more reliable, as methods have improved and allowed for objective measurements of perceptual intensities of the auditory sensation. As stated previously, one purpose of this project was to maximize the perceptual relevance of the physical measurements in this PhD project, such that the all established prediction models would be based on an emulation of the path from headphone/loudspeaker to auditory sensation. The physical measurements were therefore obtained by reproducing musical excerpts over sound reproduction systems, recording the stimuli in the listening position using a head-and-torso simulator (HATS), and processing these recordings using auditory models.

Data basis: Generalisability of predictions

For a prediction model to be of general value for a certain system group, it is important to ensure that the perceptual space spanned by the systems, on which the model is trained, are representative of the types of systems one wish to make predictions of. Depending on the level of ambition, this may require a large number of systems. In Table 1, studies including predictive modelling of sensory descriptors are listed for loudspeakers (L) and headphones (HP). Note that relevant papers from this thesis are included for comparison. Klippels loudspeaker study [39] consisted of seven separate experiments, explaining the high number of systems, musical excerpts, and subjects. An trend in Table 1 is the smaller dataset size in studies of headphones compared to studies of loudspeakers. Since the variations within loudspeakers are much larger in terms of number and size of drivers, type of cross-over filter, cabinet volume etc., this could make sense, but does not mean that the smaller number of systems and subjects is sufficient for generalised predictions of headphone characteristics.

A prediction model's power is influenced differently depending on the experimental factors of the listening test on which it is trained, such that:

- Increasing the number of perceptually different systems, increase the generalisability of the study's conclusions.
- Increasing the number of excerpts, increases the generalisability of the systems' characterisation.
- Increasing the number of subjects, increases the general validity of the study and decreases the uncertainty of the statistical inferences, e.g. the

Year	Source	Product type	Systems	Excerpts	Subjects	Ortho. dim.	Descriptors
1990	Klippel [39]	L	45	53	94	2-3 (86%-98%)	40 (7)
1991	Gabrielsson [16]	L	18	8	16	-	7
2004	Olive [61]	L	70	4	268	5 (97 %)	1
2016	Volk et al. [Paper D]	L	11	2	10	3 (89%)	6
2006	Opitz [63]	HP	4	2	12	3 (84%)	3
2010	Chon & Sung [6]	HP	8	5	10	-	1
2012	Olive & Welti [60]	HP	6	3	10	-	19
2016	Volk et al. [Paper C]	HP	8	4	18	4 (92%)	6

Table 1: Summary of audio studies including prediction modelling of sensory descriptors or preference. HP denotes headphones/earphones and L denotes loudspeakers. Not reported details are marked with '-'. Note that the Klippel study included seven separate experiments. Seven descriptors were modelled, using different subsets of the data.

confidence intervals of the mean[2].

While the above bullet list describes how to increase the generalisability of a study, it is difficult to assess the absolute generalisability of a listening test due to the vast amount of sound reproduction system variations and configurations available. One way to handle this dilemma, is to have a focused scope with regards to the prediction model. In Paper C, focus was of predictions of prototypes from one brand, and in Paper D, focus was on *affordable* 2-way dynamic compact loudspeakers. Within these narrow scopes, 8-11 systems may provide insight of some general value, while e.g. results of the referenced headphone studies with only 4-8 systems are not likely to represent headphones in general.

In Paper B, 21 headphones were included in an multi-dimensional scaling (MDS) study of dominating headphones characteristics in a effort to obtain some generalisability. The headphones were chosen to represent products currently on the market in terms of distribution between open and closed headphones as well as a selection of popular brands. Within the MDS methodology, a mathematical rule-of-thumb [40] states that the number of stimuli N needed to statistically uncover N_{dim} dimensions can be estimated by this equation: $N_{dim} = \frac{(N-1)}{4}$, e.g. 13 systems for 3 dimensions, 17 for 4, 21

[2]For subjective studies of preference, the confidence intervals may increase with the addition of subjects, if the subjects disagrees on preference, thus representing several clusters of preference. This is not the case with perceptual measurements where intensity is measured (when using normal-hearing listeners).

for 5, etc. The means, even if a set of stimuli spans X perceptual dimensions, they might not all be uncovered in the studies if less stimuli than stated by this rule-of-thumb are included in a study.

Data basis: Listener reference

In terms of prediction modelling, a fundamental issue, is how listeners perceive stimuli. When evaluating e.g. the bass strength of a number of loudspeakers, how do they determine whether a certain stimulus has a little bass, neutral bass, or a lot? Somehow, they need a reference. In models for codec testing, such as P.OLQA for telecommunication sound quality [30] and QESTRAL for spatial quality [11], the reference is simply the original musical excerpt played back over e.g. headphones. This is possible, because the degraded stimuli is played back over the same reproduction systems, and thus only the degradations due to the codec is assumed to affect listener evaluations. For evaluation of physical products it is, however, in general not possible to present the original stimulus. In [39] Klippel assumes that listeners know the recorded reference and are able to use this as an "internal reference" for assessment of systems by evaluating the deviations from this reference. This assumption is not likely to hold, as listeners have no possibility of knowing the reference: They only have the presented stimuli available, e.g. musical excerpts influenced by each of of the sound reproduction systems under test; And while they might have had heard the excerpts before, they have never heard it without influence from reproduction equipment. Consequently, in this project, it was instead assumed that listeners evaluate *"the perceived changes to the envisioned original sound"* [Paper D], with the envisioned original sound being created by listening to the test systems in the given environment (listening room). Put another way, listeners compare similarities of the presented stimuli and formed estimates of the original sound of the excerpts. To provide listeners with the opportunity to form estimates of the envisioned originals, each listening test in this project started with a familiarisation session including all or a representative selection of stimuli. The influence of this new definition of the "internal reference" was a modelling strategy were a reference was estimated by averaging over the ratings of all stimuli per musical excerpt. This question of listener reference was discussed in Paper D.

One assumption, which neither of the previous studies, nor the papers included in this thesis discusses is the fact that perceptual evaluations made by listeners have been found not to be absolute, which is a challenge in terms of making absolute predictions. This aspect have been discussed in detail in the Section 4.1, along with possible actions for dealing with this fundamental challenge.

1.2 Overview of thesis papers

The collection of papers comprising the main body of this thesis consist of two conference papers, two journal papers, and one research report. The first paper (Paper A) is a literature study, and discussed five aspects of how to obtain objective and relevant data from listening tests. The second and third paper focuses on headphones. One (Paper B) investigates the dominating perceptual differences among a large and varied set of headphones and proposes metrics characterising these differences, while the third paper (Paper C) describes metrics for prediction of well-defined sensory characteristics of a selection of prototype headphones. This paper also investigates an experimental method for validation of the prediction models when dealing with a small number of systems. The research report (Appendix in Part III) and the fourth paper (Paper D) deals with sound reproduction of loudspeakers. The report describes an early attempt of generating data for modelling, both perceptual and physical, based on evaluation of a set of sound sources generated by a baffle of loudspeakers (see Fig. I, p. 136). Unfortunately, this experiment suffered from two significant issues with listening test biases and did not have the data quality for reliable modelling. The fourth papers describes metrics for prediction of sensory characteristics of loudspeakers in a classical stereo setup. It is the culmination of both this PhD thesis and the parallel activities at SenseLab in terms of modelling prospects.

In the following chapters, relevant topics are presented and discussed in relation to the studies described in the four papers and the research report. References are included regularly to make the reader aware of how topics are linked to the relevant papers.

2 Auditory processing

While many of the traditional measurements of a loudspeaker's acoustical output are linear, human perception of this output is far from linear. The extensive research in psychophysics and psychoacoustics have revealed non-linearities in every step of the auditory processing. How a sound is perceived depends on sound level, frequency (range and sparsity), temporal characteristics, and higher level (cognitive) processes. Furthermore, differences between the sound reaching the two ears affect perception. As a result, any linear measurements of sound reproduction will be related to auditory perception in a highly indirect and incomprehensible fashion. Consequently, characterisation of loudspeakers and headphones have been largely limited to that of investigating the similarity between physical measurements and the original test signals (see e.g. [15, 81]) or musical excerpts as they appear on the medium (CD). It is, however, of further value to know how deviations from perfect/preferred reproduction are perceived, and how the sum of these

deviations affect the dominating characteristics of the auditory sensation of the reproduction (or even the emotional consequences).

In this project it was believed that metrics having causal relations between the physical and the perceptual domains would lead to be the most robust outcome. In practical terms, this meant that prior knowledge about auditory perception was part of every step of the modelling process.

To obtain metrics with a direct relationship to listeners' auditory perception, the prediction models were based on perceptual ratings from listening tests and a physical representation of the sound reaching the ears of the listeners. The physical representation consisted of recordings of the sound reproduced by the systems included in the study captured by a head-and-torso simulator. These recordings were processed using an auditory model able to account for the non-linearities of human auditory processing. A description of the loudness model, favoured in the studies of this thesis, is present in the next section in terms of advantages and limitations for this specific purpose.

2.1 Loudness perception

The main auditory model used in this project, was the time-varying (TV) loudness model by Glasberg & Moore [20]. The model is based on their stationary loudness model [55], which took a frequency spectrum as an input. This spectrum represented the average of both the left- and right ear signal and the full duration of the stimulus. The TV loudness model added two temporal aspects: 1) Calculation of instantaneous loudness for every millisecond and 2) Calculation of a short- and long term loudness. The first addition ensures calculation of masking effects as a function of time, while the second addition accounts for temporal integration of loudness, i.e. that a stimulus with a short duration (transients) are perceived to have a lower loudness level, than a stimulus of identical sound intensity, but a longer duration.

This loudness model can account for observed non-linearities of the auditory system as a function of sound levels, frequency, bandwidth, and temporal aspects such at temporal integration, amplitude modulation and onset/offset perception. A diagram of the loudness model is depicted in Fig. 3. Note that the calculation of overall specific loudness was not part of the original model, but was the basis of the loudness calculations for the metrics proposed in this thesis. Extraction and calculation of the overall specific loudness thereby bypassed the temporal integrators of the original TV loudness model. These temporal integrators, short- and long-term, relies on automatic gain control with time-constants that were tuned to fit data on perception of stimuli of various durations. However, no consensus currently exists regarding the type and number of these temporal integrators and their time-constants [20, 25, 58]. Hots et al. [25] concluded on the basis of a proposal by Kumagai et al. [41] and data from Poulsen [68] that two integrators

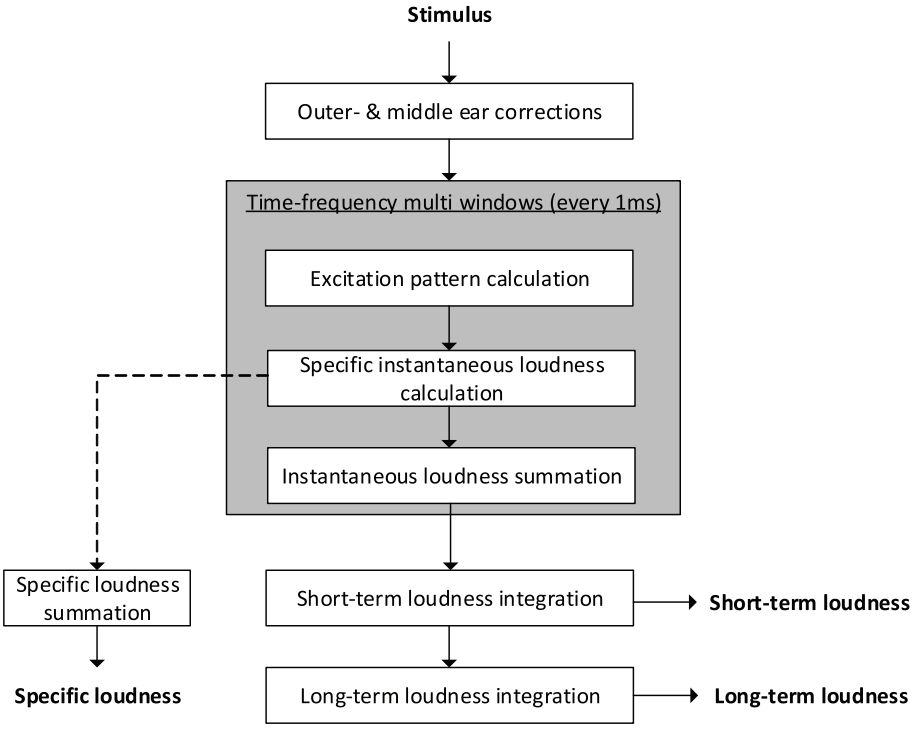


Fig. 3: Diagram of the time-varying loudness model Glasberg & Moore [20]. A spectral loudness representatation is calculated for every millisecond. Note: The calculation of (overall) specific loudness is not part of the original model. Diagram adapted from [55] and [19].

in parallel may better predict current loudness experiment data compared to the serial approach of Glasberg and Moore (depicted in Figure 3). Both approaches, however, perform a spectral integration prior to the temporal integration, which is unwanted for the modelling approach in this thesis. In [58] it is speculated that loudness is calculated by multiple approaches in the auditory system on the basis of a spectro-temporal excitation pattern (STEP), which supports the idea behind the approach taken in this PhD project. Unfortunately, the integrator time-constants as a function of both spectrum and duration is not presently know and consequently, calculation of the overall specific loudness was estimated using simple averaging over the time dimension. For stimuli with large variations over time, simple averaging may not be representative of a listener's perceived basis for evaluation. An example of a large variation could be the start of a dominant bass drum, the start of an instrument solo, large level changes (start of a chorus or a 'bridge' segment) etc. The musical excerpts of the studies of this thesis were all cut, such that they had limited variation over time. This was done for reasons of good audio perceptual evaluation practice: 1) To have all assessors base their evaluations on the stimuli as a whole and not smaller parts, e.g. either the start or end, and 2) To ensure a looping of the excerpts, which does not remove listeners' focus from the evaluation task.

The TV loudness model includes simplified binaural summation of loudness. The binaural summation processing steps were however revised in a updated loudness model for stationary signal [54], which modelled contralateral binaural inhibition, i.e. the inhibition of loudness in the left ear caused sound reaching the right ear and visa versa. Generally, it has been found that binaural summation of identical left and right stimuli lead to an increase in loudness corresponding to 5-6 dB (a loudness factor of 1.4-1.5[3]), while a 10 dB increase (a loudness factor of 2) would be anticipated with perfect loudness summation. The degree of inhibition is, however, dependent on the level difference between ears and possibly the overall level. Furthermore, the degree of inhibition depends on frequency content; at least for low frequencies, where phase information is conveyed accurately from the cochlears to the central auditory system (see e.g. [35]) responsible for comparing and processing signals from both left- and right ear. In a summary article from 2014 [53], Moore described a revised time-varying model including the same binaural inhibition process as proposed in [54], which is currently being standardized in ISO/DIS 532-2. This binaural time-varying loudness model (BTV) was kindly provided by Glasberg and Moore for use in the current PhD project and used in the study described in Paper D. Here, the prediction models were trained using either the TV- or the BTV loudness model. The result was

[3]The conversion between loudness factor and difference in SPL is described by $\Delta L = 10\log_2(z)$, where z is the loudness factor.

identical or very similar metrics in terms of frequency ranges, but the TV model led to better correlations in almost all cases. Consequently, the results of the BTV were not reported in Paper D, due to the lack of improvement over the simpler TV model and the lack of published articles documenting its performance. The BTV loudness model may, however, be of potential additional value compared to the TV loudness model in experiments with other musical excerpts or types of stimuli, as excerpts included in the experiment of Paper D were similar in terms of sound level between channels, i.e. had a symmetrical sound image.

Loudness model representativeness

The use of a loudness model to predict perceptual ratings relies on the assumption that the loudness model is representative of the listeners making the perceptual ratings. This assumption relies on another assumption, namely that both are representative in terms of hearing capabilities (characteristics) of an average normal hearing listener. Obtaining model coefficients and perceptual ratings meeting this requirement would require a large group of listeners and extensive validation. An alternative approach (not investigated in this PhD project) was utilized by Jepsen et al. [33], who made measurements of the hearing capabilities of ten individuals with sensorineural hearing loss and fitted an advanced auditory model [34] to the individual. This approach have the potential, not only to improve loudness predictions, but also to gain metrics that represents the true auditory processing of listeners more closely.

Another issue with current loudness models is that they are rarely tested with realistic stimuli as is the input in the prediction models of this project. Doing so could lead to better loudness models or better fitting coefficients in the current ones. One study, showing the need for this, by Soulodre [76], found that in a comparison of ten (unnamed) loudness models' ability to predict perceptual ratings of loudness of eight programme materials with either speech, environmental sounds, music or a combination, the best loudness predictor was $L_{EQ}(RLB)$. $L_{EQ}(RLB)$ is simply the average sound pressure level weighted by a frequency curve, RLB, which is identical to the B-weighting curve at low frequencies and flat at high frequencies. I.e. a "primitive" predictor without emulation of human hearing in regards to temporal processing, compression, binaural processing, etc. In general perceptually-driven prediction models has however been found to be more robust than "technical" measures [23].

3 Perceptual evaluation

With regards to evaluation of sound quality and sound character, loudspeaker and headphone manufacturers have historically always relied on human auditory perception; Either for iterative evaluation as part of the development process or for confirmatory evaluation as a supplement for measurements of technical aspects of the reproduction (frequency response, impedance, sensitivity, etc.). With pioneers like Alf Gabrielsson, Floyd E. Toole, and others, the systematic use of listeners for perceptual evaluations became a recognized method of performance evaluation and is now considered the most relevant technique for audio evaluation. Some debate, however, still exists regarding the nature of perceptual evaluations and in most fora perceptual evaluations are referred to as subjective evaluations, as opposed to objective evaluations. Experts within the domain of perceptual audio evaluations, however, divide perceptual evaluations into two categories of which one is objective and one is subjective (see e.g. [1, p. 3]):

- Perceptual measurements
- Affective measurements

Perceptual measurements are objective evaluations of system characteristics, e.g. bass strength, while affective measurements are subjective expressions of preference or sense of quality. A discussion of the objectivity of perceptual measurements is included in Paper A based on definitions by Jens Blauert. The concept that listeners are able to make both objective and subjective evaluations is founded on the filter model framework [64], shown in Fig. 4. The

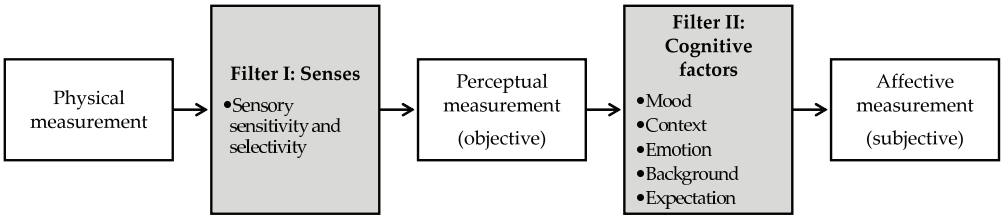


Fig. 4: The filter model. Perceptual measurements are based on the sensory acuity only, while affective measured are also affected by higher cognitive factors, such as prior experiences, mood, and emotions. Diagram adapted from [1].

perceptual measurements are assumed only to depend on the sensory acuity of the auditory system and thus measurements with listeners in one place can be reproduced by listeners in another place in terms of statistical inferences. This is not (necessarily) the case with affective measurements. Some may prefer more bass, others more treble and their preference may change from one day to the next due to influences such as context, mood etc. Note

that all modelling effort in this PhD project, dealt solely with prediction of objective perceptual measurements of product characteristics.

3.1 Sensory descriptors

The characteristics of sound reproduction systems, which was sought predicted in this project, consisted of sensory descriptors from a list of descriptors developed and organised in a Sound wheel [66]. The list is designed to perform a characterisation of sound reproduction systems, i.e. the most common senory descriptors related to the perception of the changes to the envisioned original sound. The purpose of the Sound wheel was three-fold: 1) To make the sensory descriptor elicitation process more efficient by selecting descriptors from an established set of descriptors, and 2) to potentially reduce noise in the data, by having listeners use the same descriptors continuously to improve their understanding of the descriptors and additionally improve consensus among listeners, 3) To facilitate communication across diverse audiences such as product developers, manufacturers, sensory experts, retail business and consumers. Furthermore, having a closed set of descriptors allow systematic training. A challenge with many of the elicitation methods available (comprehensive review in [8]) is that selecting sensory descriptors and making proper definitions is a demanding task requiring understanding of the desired properties of sensory descriptors [1, 66].

The present version of the Sound wheel [84] with 41 sensory descriptors (+ four sub descriptors) is depicted in Fig. 5[4]. The sensory descriptors modelled during this project are summarized in Table 2, page 34 and comprises eight descriptors from the Sound wheel. These were selected in a two step process: Firstly, during consensus elicitation meetings with listeners for the headphones or loudspeakers of each study and secondly, by only keeping descriptors whos ratings met a number of data quality requirements (see details in Paper C, Section 2.5). Most of the sensory descriptors (6) were directly related to spectral characteristics, because the output of the consensus elicitations mainly comprised spectral aspects and because listeners performed well in terms of discrimination and reliability (calculated using eGauge [49]), i.e. few spectral sensory descriptors were removed during data quality investigations. This dominance of spectral characteristics is typical in perceptual audio evaluations, as previously mentioned and described in Paper A, Section 3.

[4]The inner most circle in [84], Basic Audio Quality, was removed here to clearly separate sensory descriptors describing perceptual measurements from descriptors describing affective measurements.

Fig. 5: Sound wheel of sensory descriptors for perceptual characterisation of reproduced audio. The rings (in→out) reprensents, groups, subgroups, and sensory descriptors respectively. Adapted from [66, 84] and generated using an Aculocity visualization tool (`www.aculocity.com/labs/sunburst-chart`).

3.2 Independence between sensory descriptors

In Paper A, the choice of descriptors was also discussed in terms of independence. Indepedence is described as a desired properties of descriptors [1, 66], but a literature study showed a discrepancy between the number of sensory descriptors (large) in some studies and the number of independent statistical dimensions (small) in other studies. Large dependencies between sensory descriptors are (viewed as) inefficient and a sign that fewer/better descriptors could have been used. This makes it of interest to measure the dominating perceptual dimensions within each product category of sound reproduction system, as it provides an overview of which categories of sensory descriptors are of most importance. This was the topic of Paper B, where the dominating perceptual difference between 21 headphones were investigated using a pair-wise comparison scheme with evaluation of the degree of dissimilarity between pairs rated on a continuous rating scale and analysed using MDS.

4 Modelling and validating

When modelling the link between the physical output of for example a loudspeaker and the rated perceptual intensity of a sensory attribute, the nature of the two measurement methods must be considered. Physical measurements are absolute measures with well-defined units. Given that e.g. a sound level meter is calibrated correctly, an isolated and stable sound source will measure the same every time (within its specified accuracy) and measurements of other sound sources will not be affected by the first measurement, as they are independent. An intensity rating of a sensory descriptor has been found not to be absolute. In a study by Brockhoff [4], significant differences in scale usage (variance heterogeneity) was found in 91 % of the investigated studies.

For the purpose of predicting perceptual characteristics, it was of interest to understand how to deal with the challenge of these context effects on listeners ratings; Specifically, how to possibly minimize the influence of context on ratings and how to predict ratings that relies on context. These questions are described in detail in the next subsections, followed by a description of the proposed prediction models, an introduction to validation techniques, and finally an introduction to the statistical metrics used for evaluation of model performance.

4.1 Handling relative ratings in prediction modelling

Background

Although perceptual ratings can be considered objective (see definition and discussion in Paper A), they are dependent on a number of variables in the

setup of the listening test [1]: The range and number of systems included in the test setup, the (musical) excerpts selected for the evaluation, the sensitivity of the listeners, the scale used, the verbal anchors, etc. The consequence of these variables, in terms of rating scale usage, are generally divided into two parts: a shift and a scaling effect. Due to reasons of simplicity, these effects are often assumed to be linear or approximately linear. For a shift, an additive effect, all system ratings are moved up or down on the scale, while the distance between ratings are kept constant. With regards to the scaling effect, a multiplicative term, perceptual ratings are moved closer together or further apart, while the ratio between them are unaffected. These two effects can be described by a general linear transformation: $x' = a \cdot x + b$. In [86] another type of bias is described, which causes non-linear changes in sensory ratings: "Bias due to perceptually non-linear scale". If the assessment scale has three or more labels, e.g. verbal, which listeners perceive as being positioned with non-linear distances, they may adjust their ratings to fit the imagined scale, which may even differ from listener to listener. This would affect the accuracy of the perceptual measurements, specifically the statistical inferences of e.g. mean ratings, but thereby becomes a problem from a modelling perspective. To avoid this potential bias in the current project, all the rating scales used were limited to two verbal anchors near the end-points (extremes) of the scales.

In the acoustics and audio journals, the effects causing sensory attribute ratings to be relative due to the choices in the listening experiment setup are well-described (see e.g. an overview in [1]). The main concern regarding listeners' scale usage, with regards to predictive modelling, is the context effect. One type of context effect is the Range Equalizing Bias [86], [69, pp. 207-233], which occurs when listeners adapt to the range of intensity for the given set of stimuli. Due to the adaptation, they tend to use the entirety of the scale for the ratings somewhat independent of the stimuli intensity range. Another context effect is the Centering bias [86], [69, pp. 105-120], where listeners tend to centre the ratings symmetrically around the middle of the scale. These effects can be reduced, but not eliminated, by including verbal anchors etc.

The influences of the context effect and other effects leading to relative listener scale usage are not taken into account in the existing prediction modelling efforts; Neither in the ITU recommendations (PEAQ [27], POLQA [30]), nor the conference papers or journal articles listed in Fig. 2. The modelling approach ranges from simple visual inspection to correlation analysis, analysis of variance (ANOVA), linear regression analysis, principal component analysis (PCA), Factor analysis, and MDS analysis as well as use of auditory models, but none considers this issue.

The Range Equalizing bias was likely observed in the one of the loudspeaker experiments conducted in this project (see Appendix III), as is clear

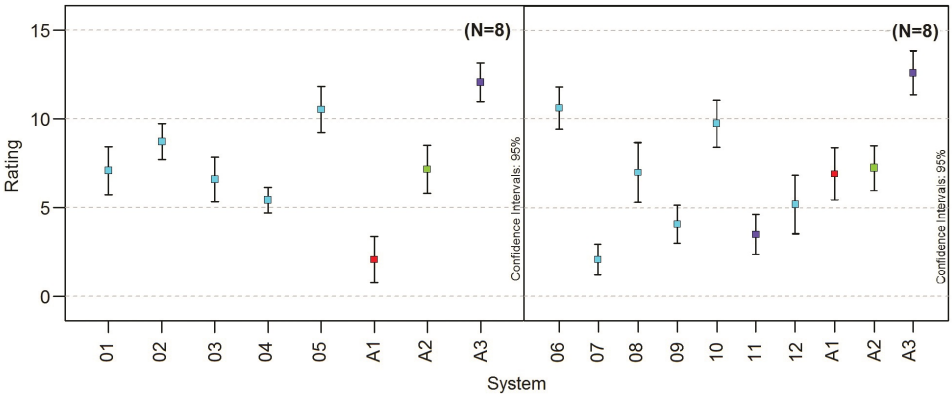


Fig. 6: Example of how the context (range of systems) can affect listeners' scale usage in listening tests. The left and right plots depicted results from two identical design of experiment, but with different sets of loudspeakers (see Appendix III). The plots depicts the mean ratings of eight listeners with 95 % confidence intervals. Systems 'A1'–'A3', were included in both.

from the example with the sensory descriptor Bass strength in Figure 6. In this example all experimental variables were kept constant from Test 1 (left) to Test 2 (right), except the set of loudspeakers included; meaning that the setup, the sensory attributes evaluated, and the listeners were identical. Three anchor loudspeakers were, however, included in both Test 1 and Test 2, which made it possible to compare the ratings across the otherwise identical tests. From Figure 6 it is evident that A1 was rated significantly different between Test 1 and Test 2. Consequently, in these two tests, context effects could not have been ignored if accuracy was to be achieved in a prediction model trained with these data. In [69] it is advised to use an indirect method of magnitude estimation to avoid Range Equalization biases, which would be a challenge to use in all cases due to demands for standardized test methods, time constraints regarding listeners' participant time etc. It could, however, potentially be a beneficial methodology for collection of training or validation data for the prediction models.

Strategies for handling relative scale usage

A number of possible actions could be considered, either to fix scale usage, or deal with a relative scale usage:

1. Including anchor loudspeakers or auditory references (at e.g. a low-, mid-, and high points on the scale)
2. Using (multivariate) data aggregation techniques
3. Intensive training of listeners in consensus scale usage for each sensory

attribute

4. Making product type specific models - valid only for a subset of audio reproduction systems.

A fifth solution could be to estimate the effects of all experimental variables and attempt to compensate for these. This is, however, not presently a viable solution, as many of the effects and the interactions between them are not yet fully understood. The four listed strategies are discussed in the following paragraphs. All of the listed actions were taken or tested during this project. The first two action points were tested in the research study described in the Appendix in Part III. Since both tests have biases in the experimental design, the data aggregation technique is, however, not described.

1. Anchor loudspeakers or auditory references
To limit the influence of the range equalization bias, a number of extra system, e.g. loudspeakers, can be included, which should ideally span the perceptual range of the set of systems for all included sensory descriptors. Typically, three are used, one at each extreme and one in the centre. These extra loudspeakers provides reference points for listeners performing multiple stimulus tests, thereby anchoring the scale usage. These anchor loudspeakers must be included in all tests, desired to be compared or modelled. An inherent challenge of this approach is choosing the anchor loudspeakers. In the scenario where the anchor loudspeakers are very different from the remaining loudspeakers, the ratings of these may be squeezed together causing an unwanted reduction in rating resolution. As the anchor loudspeakers must be fixed for all the tests that it is desired to compare, choosing suited anchor loudspeakers can become a limiting experimental factor. Furthermore, the introduction of extra loudspeakers, may limit the number of loudspeaker possible to include in the test, as the listener's comparison task becomes exponentially more complex (in terms of pairwise comparisons) with the number of loudspeakers included. This can to some degree be resolved by including anchor auditory references instead of physical loudspeakers. These should correspond to a fixed pre-defined rating for each sensory attribute, and would thereby not require evaluation by the listeners. Finding good auditory anchor references is, however, a complex task, which must be repeated for every sensory descriptor (and from a rigid point of view also for every change of descriptor definition).

2. Data aggregation
In the case where datasets with loudspeaker evaluations are available that included anchor loudspeakers, it is possible to use data aggregation techniques to correct for shift and scaling effects. This is relevant if the anchor

loudspeakers got different mean values with respect to the statistical inference. While this technique has been used in the Sensory food industry for some time, few papers are published on the subject investigating these techniques in depth. One concern is that the uncertainties of the dataset being transformed to the scale of the target dataset are amplified/altered during the transformation. Another concern is the need to make an assumption of the type of transformation (e.g. linear, log-linear etc.), which may change the relationship between the loudspeaker ratings in a non-optimum fashion and thereby reduce the relationship between ratings and listener perception. The data aggregation can be done either by a unidimensional linear method for the individual sensory attribute ratings or by a multivariate approach, such as the Generalized Procrustes analysis (GPA) [21]. For the unidimensional linear approach a minimum of one anchor system is needed for correcting biases related to shifts, while a minimum of two anchors systems are needed for correcting biases related to scaling effects.

3. Intensive training of listeners in consensus scale usage
For most perceptual evaluations of system characteristics, a panel of listeners is trained to become experienced or experts in evaluation of the specific sensory descriptors used within their field of expertise. During this training listeners are trained in the definition of the sensory descriptors, as well as in rating scale usage, such that consensus is attempted with regards to the rating of a sensory descriptor with a given intensity. Brockhoff et al. [3] recently analysed a large body of sensory data (from SensoBase) where experienced of expert listeners were used, to investigate to which degree significant interaction effects between listeners and systems were found and how often these could be explained by shifts and scaling effects. His study showed that in studies where significant interactions between listeners and products were found, significant differences between listeners could be explained by scaling effects in 57.2 % of the cases. Counting on listener training as a mean to obtain stable sensory ratings are therefore generally not considered a complete solution on its own to solve the challenge of relative sensory ratings.

4. Specific modelling based on product type
The range Range Equalizing bias cannot be avoided completely, which leads to an important realisation with regards to the prediction models: A prediction model may be needed for each product type where the rating scales are used differently. An example could be the ratings of Bass depth in small cheap loudspeakers maybe depend on a different part of the frequency response, than evaluation of Bass depth in high-end floor-standing loudspeakers. The definition of a product type in this context is not clear cut. For headphones, it might constitute two groups: closed-back and open-back, but a split into groups could also be done on the basis of overall sound quality,

low-end vs. high-end. The points is, not to assume that ratings of one sensory descriptor can be modelled with only one prediction model valid for all sound reproduction systems. Instead the prediction power of a model, i.e. gauged by a correlation analysis, could be used to determine whether products belong to the same perceptual product group, or whether they should be separated into multiple groups, e.g. by performing cluster analysis.

4.2 Overview of the proposed prediction models

The ambition behind the structure of the proposed prediction models of this project, was to approximate the human auditory processing leading to the perception of sound reproduction system characteristics to a large extend, while simultaneously limiting the complexity of the models. As a result, simple modelling approaches were initially attempted, and with good results. Two main modelling frameworks were established: One generic framework for prediction of sensory descriptors related to timbrel aspects, and a second framework established specifically for prediction of the Dark-Bright descriptor. Dark-Bright is defined as the perception of the spectral balance of the sound reproduction. The first framework is first introduced in Paper B, while the second framework was introduced in Paper C and refined in Paper D.

Generic timbre prediction framework
The first prediction modelling framework is described by Eq. 1, where $Dens_m(f)$ denotes the temporal mean of the instantaneous specific loudness, while A–D denotes the frequency limits, which the specific loudness is summed over.

$$metric = \frac{\text{AB range}}{\text{CD range}} = \frac{\sum_{f=A}^{B} Dens_m(f)}{\sum_{f=C}^{D} Dens_m(f)} \tag{1}$$

This equation thereby constitutes a loudness ratio between two frequency ranges of specific loudness. This allows estimation of the relative loudness of a given AB frequency range, either in relation to the full frequency range or a more narrow frequency range CD. The two ranges were found in two steps. In step one, the Pearson correlation between the perceptual ratings for *a* musical excerpt and all possible combinations of AB- and CD ranges were calculated (with some constraints on minimum range and range overlap specified in Paper B). This led to a correlation grid/matrix with a resolution of either 0.1 or 0.25 critical bands, depending on the utilized loudness model. An example of such a correlation grid is depicted in Fig. 7, with the

frequency limit A on the horizontal axis, B on the vertical axis, and CD set to the full range. Any AB-range leads to a different correlation with the perceptual ratings. Note that grid points with correlation coefficients $|r| < 0.8$ are represented in gray to improve the overview. In this example a AB range of 20-210 Hz led to the highest correlation, as well as the complimentary range 210-15000 Hz. In a second step, the correlation matrices for each musical

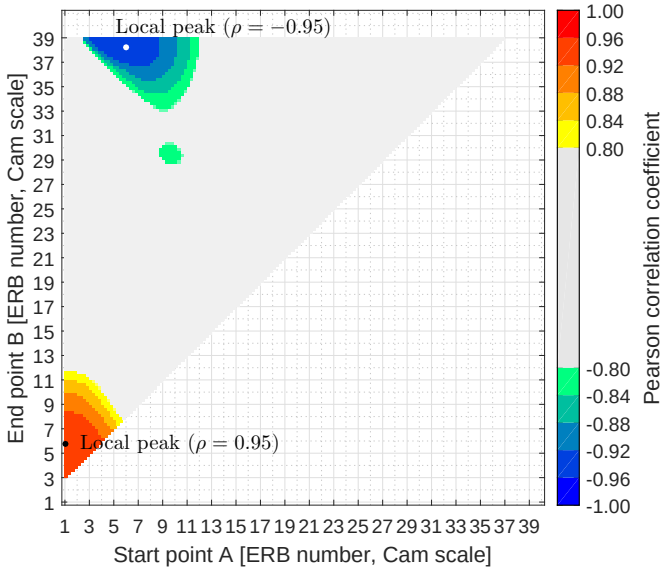


Fig. 7: Example of a correlation grid for the generic prediction model framework. Plotted as a function of the start frequency A and end frequency B (with the CD range fixed to the full specific loudness range in this example). The lightest grey represents all values within $r = \pm 0.8$. The figure is a reprint of Fig. 7 in Paper B.

excerpt were first averaged across and then the frequency limits, A–D were chosen as the limits where the highest absolute Pearson correlation, $|r|$, was located. This was done to obtain AB- and CD loudness ranges with limited dependence on a single musical excerpt. The performance of this framework were tested in two different versions in Paper B: One, with both AB- and CD being optimised to obtain the maximum correlation with perceptual data and one with the CD-range fixed to the full specific loudness range (20-15000 Hz). These two versions (mostly resulting in different AB- and CD ranges) led to almost identical levels of correlations. In the subsequent two papers, only results from the version with the fixed CD range were reported, as the same trends was found. Furthermore, the risk of obtaining non-meaningful frequency ranges increased with the added complexity of the variable CD range. As an example, a prediction model of midrange strength, having both

AB and CD ranges in the treble range in likely not be a robust predictor of sound reproduction systems in general, due to lack of causality. As a result, such a prediction model is unlikely to perform well in prediction of new headphones or loudspeakers. For some sensory descriptors, such as Clean, it may not be possible to define meaningful frequency ranges.

Dark-Bright prediction

Since the sensory descriptor Dark-Bright is defined as the spectral balance of the sound reproduction, the generic framework was not deemed appropriate, although it could provide reasonable predictions, when the CD range was not fixed. Another approach was established that lead to better predictions. This approach was based on finding the spectral centroid, i.e. the point of equal energy between the lower frequency content and the upper frequency content. This had been used before in literature, but only on linear frequency spectra and not on specific loudness spectra, i.e. processed through a loudness model prior to finding the centroid. The problem of finding the specific loudness centroid, here referred to as the perceptual centroid, was first described in Paper C, Eq. C.5 and reprinted here in Eq. 2. $Dens_m(f)$ is the temporal mean of the instantaneous specific loudness, and f_{MIN}, f_{CEN}, f_{MAX} are the minimum, centroid, and maximum centre frequencies respectively.

$$\begin{aligned} \min_{b_{CEN}\in\mathbb{Z}} \quad & \left| \sum_{b=b_{MIN}}^{b_{CEN}} Dens_m(b) - \sum_{b=b_{CEN}+1}^{b_{MAX}} Dens_m(b) \right| \\ \text{subject to} \quad & b_{MIN} \geq b_{CEN} \leq b_{MAX} \end{aligned} \tag{2}$$

In Paper D a revised version of the Dark-Bright prediction model was suggested, which were based on the assumption that the midrange frequencies might not have the same influence on perception of spectral balance as frequencies at the bass- and treble range. A proposed addition was to add a weighting function prior to finding the perceptual centroid. One simple weighting function was tested: An upside-down rectangular window, depicted in Fig. 8. A percentage weighting coefficient, p, was optimised such that the highest possible correlation with the perceptual ratings of Dark-Bright was obtained. The rectangular window improved correlation with the perceptual ratings compared with the originally proposed prediction model. This weighting function was not found likely to be part of human auditory processing due to the discontinuity in weighting coefficients at the start and end of the upside-down rectangular window, but indicated that the existence of a weighting function is likely.

Both of these prediction model frameworks were based on current knowledge about the human auditory processing of sound and tested on perceptual evaluation of both headphones and loudspeakers. While one methodology

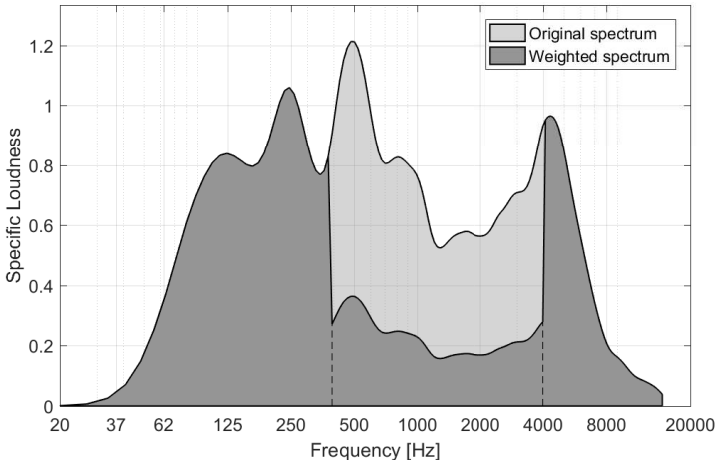


Fig. 8: Illustration of a midrange weighting function for prediction of Dark-Bright. An upside-down rectangular window with a weigthing coefficient of $p = 30\,\%$ is multiplied with the specific loudness spectra, prior to calculation of the perceptual centroid.

was specifically designed for modelling of Dark-Bright, the other has potential for modelling of additional spectral-related sensory descriptors from the Sound wheel [66], such as Canny, Boomy, Boxy, and Full.

4.3 Influences of musical excerpts

One concern, when using perceptual evaluations as the basis for prediction modelling, is the influence of the musical excerpts on the ratings. This influence comprises of both a shift and a scaling effect. The shift depends on the absolute intensity of the sensory descriptor under evaluation, e.g. treble strength. Sound reproduction systems may receive a lower rating when evaluated using an excerpt with little treble. The scaling effect is of more interest as the span of ratings given for e.g. treble strength may depend on the amount of treble in the excerpt. This is exemplified in Fig. 9.

In terms of obtaining a set of generally valid metrics for prediction of sensory descriptors these effects pose a challenge. Both are linear effects and can be compensated e.g. by averaging the perceptual ratings over excerpts. It might, however, be of interest to investigate whether this averaging is meaningful. If significant non-linear effects are present for some excerpts, it may not be meaningful to average over all excerpts. Additional, it might make sense to have multiple genre-dependent prediction models, if some musical excerpts/genres are highly different from the others. In the latter case, knowing the increase in intensity that is required to obtain a clear perceptual difference for a certain genre of music, would enable design targets better

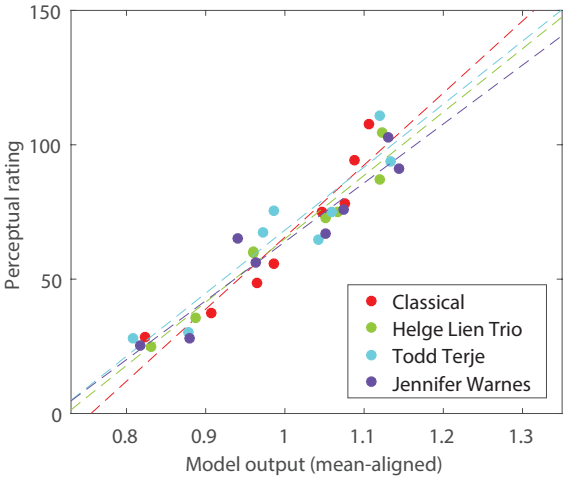


Fig. 9: Scaling dependence on musical excerpt for the sensory descriptor Treble strength. The dotted lines are the best-fit linear regressions for each musical excerpt respectively.

fitted to the target group of a sound reproduction system. In the example in Fig. 9 it is for instance evident that a smaller increase in treble loudness is needed to increase the perception of treble strength for a classical excerpt, than for the other musical excerpts (although the scaling effect is tiny in the example). It may not be a coincident that the classical excerpts differed the most, as the long-term spectral distribution for classical music has been found to be different from other genres and has greater variations within the genre [13, 22].

4.4 Validation technique

As argued in a previous section, validation is an important part of gauging the performance of a prediction model. The point of validation is to test whether a model's predictions are general and representative for a target group of systems, e.g. compact loudspeakers, and not only the subset used for training of the model. A prediction model is optimised using statistical fitting metrics, such as the correlation coefficient, r. If the fitting metrics increase for a subset of data - the training subset - without increasing the prediction of other systems from the target group, over-fitting is occurring. Over-fitting causes overly optimistic fitting statistics, possibly over-complication of the model, and/or a decrease of prediction performance within the target group systems as a whole.

In this project, two validation techniques were considered:

1. Splitting of data into a training and a validation subset, and

2. Cross-validation

Item 1 in the list is considered the most ideal method, as the statistical fitting metrics are used only on a validation subset of the collected data, and the training subset is used only for training of the prediction model. This method was utilised in Paper D with validation of prediction models for characterisation of loudspeakers in a stereo setup. A challenge with this method is, however, that it requires a large dataset, as, ideally, both the training and the validation set should be representative of the target group of systems. The method provides a clear separation of training and validation, but with an inefficient use of data and consequently a requirement for a larger dataset, than is needed with the cross-validation method. This method (see e.g. [10, Chap. 11]) uses the full dataset for both training and validation, but splits the set into k folds, of which $k-1$ are used for training and 1 is used for validation at a time. This is repeated k times with each fold being in the validation set once. Commonly k is set to 10, but the optimum number of folds depends on the size and dimensionality of the dataset. A special case of the k-fold method is the leave-one-out (LOO) strategy in which the validation set consists of data points from exactly one system. Using cross-validation allows for calculations of e.g. correlations coefficients, which better represents the full dataset. Cross-validation is widely used for smaller datasets, where the purpose of validation is to estimate how a model will perform in general on other data. As mentioned previously, validation is not widely used within audio reproduction research, as evident from the studies summarised in Fig. 2 of which few included validation. The ones that do are briefly discussed in the next paragraph.

The standardisation models developed by the International Telecommunication Union (ITU) [27, 29, 30] all use the first validation technique with strict separation between the training and validation data subsets. These models were both trained and validated on large ecologically representative datasets. The same validation method was used for the distorting model by Tan et al. [56], but using artificial signals. Some of these were virtual representations of real sound reproduction systems (mobile phones) and some were manually distorted signals. Using artifical signals for validation is not without problems, as was discovered in a validation of [56] with real-devices as part of a master project in colaboration with DELTA SenseLab [62]. Harlander et al. also performed validation of models by others, when they tested previouly proposed perceptual models using publicly available databases. The alternative method, cross-validation, was used only in one of the studies of Fig. 2, namely the QESTRAL project on spatial sound quality [31].

In Paper C another approach was utilized. Here, a bootstrap method (see e.g. [59, Chap. 16]) was used for estimation of model parameters - in addition to calculation of correlation coefficients. Bootstrapping is a method

that provides improved insight about the collected dataset, specifically the distribution of the data. It does not, however, provide information with regards of how representable the collected dataset is of the prediction model's target group of systems. The insight is achieved by repeated *sampling with replacement* from a dataset. One sampling iteration results in a 'new' dataset of the same size as the original dataset, but with a different representation of systems. If the dataset included 10 systems, a resampled dataset might have e.g. 7 systems, with three of them being represented twice. This is equivalent to adding 'weights' to each systems from $[0 - N]$, where N is the number of systems. By calculating new model parameters for each iteration of the bootstrapping (of a total of e.g. 1000 iterations), the method provides an estimated distribution of the parameter values for improvement of both the model and estimates of its performance. In Paper C the distribution of model parameters was categorised and the percentage of iterations within each category used for choosing the optimum parameters. This combination of bootstrapping and categorisation was somewhat experimental and was proposed to investigate methods of optimising models on the basis of small datasets (< 10 systems).

Gauging model performance

In gauging the performance of the prediction models of this project, it was important to consider the uncertainties of the data from the listening tests in combination with the performance metric, the correlation coefficient. Since the modelling was conducted used averages over listeners, the correlation coefficient may overestimate the prediction power of the models. The risk is illustrated in Fig. 10 with example data. The vertical axis have the perceptual rating and the horizontal axis have an example output predicted by a model. The errorbars represent the 95 % confidence intervals of the perceptual ratings. If the perceptual ratings were without any uncertainties, a perfect prediction model would predict values strictly on the diagonal. With uncertainties on the perceptual data, any optimisation of the predictions of a model beyond the point were CI's of all data points overlaps with the diagonal cannot be verified. Two competing models, that both perform at this level may, however, have different correlation coefficients. In this case it is invalid to chose one over the other without improving the data basis, i.e. by reducing the uncertainties of the perceptual data. This could for instance be achieved by supplementing the perceptual dataset with additional listener ratings, or by repeating the listening test with more careful control of the experimental variables - if possible.

Another issue with the correlation coefficient is that shift and scaling effects (discussed in Section 4.1) have no influence on the Pearson correlation coefficient, meaning that other performance metrics, such as the RMSE is

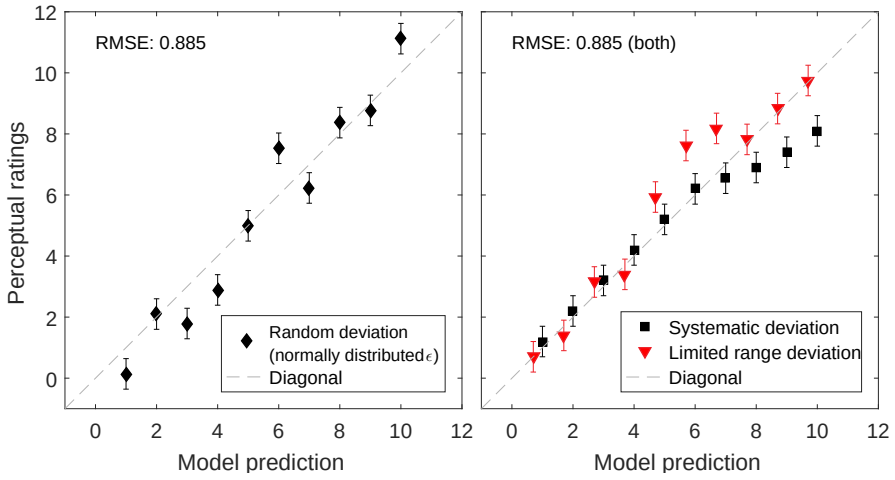


Fig. 10: Three examples of prediction model outputs - all have the same correlation coefficient and RMSE. The left subplot depicts a model with errors, which are normally distributed. The right subplot depicts models with systematic errors at the upper- or middle rating scale range respectively.

needed as well to gauge the performance with validation data sets. RMSE is a measure of the average deviation from the ideal linear relationship, i.e. the diagonals of Fig. 10. One thing must be kept in mind, when using the RMSE as a performance metric: competing models with identical coefficients may not be equally suited. This is also illustrated in Fig. 10, where the output of a model in the left panel has errors randomly distributed around the diagonal, while the two models in the right panel both have systematic errors in the mid- and high range of the rating scale respectively. Consequently, the left panel model, may be the best choice despite the three models having identical RMSE. This emphasizes the need of using visual inspection as part of the process of gauging prediction performance. Visual inspection is also useful for getting an indicating of whether the Pearson correlation formula or another correlation formula, Spearman or Kendall, is best suited.

5 Summary of findings

5.1 Main results and contributions

The main outcome of this project was a total of 12 metrics. These are summarised in Table 2. The terms 'AB-range' and 'CD-range' are defined in Secion 4.2. The metrics provide the means for 1) predictions of the dominating characteristics (defined by othogonal MDS dimensions) differentiating head-

phones, as well as prediction of perceptual characteristics of 2) headphones and 3) loudspeakers. The latter two defined by sensory descriptors in the Sound wheel [66]. All proposed metrics describe characteristics related to spectral aspect of sound reproduction systems. This was not intentionally so, but a results of listeners' performing better in evaluations of spectral characteristics.

All of the metrics performed at a high level ($r \geq 0.76$), with Pearson correlation coefficients for eight of them being above $r \geq 0.87$, e.g. able to explain or predict $\geq 75\,\%$ of the variation of the perceptual ratings. While the metrics for the prototype headphones of Paper C, may not be of general use, the remaining metrics are likely valid for prediction of headphones and compact loudspeaker characteristics in general.

Metric	Paper	Type	AB-range	CD-range	$\lvert r \rvert$
DeepBass	B	HP	20-190	20-15000	0.95
Bass/Mid	B	HP	20-220	246-3300	0.95
Bass+Mid	B	HP	140-2000	20-15000	0.93
Mid/Treble	B	HP	330-1400	3700-14000	0.92
Bass strength*	C	HP	20-15000	210-15000	0.79
Midrange strength	C	HP	690-15000	20-15000	0.97
Treble strength	C	HP	8700-15000	20-15000	0.99
Clean	C	HP	20-8200	20-15000	0.76
Dark-Bright	C	HP	-	-	0.95
BassPunch	D	L	20-72	20-15000	0.90**
Brilliance	D	L	8300-10000	20-15000	0.96
Dark-Bright (R)	D	L	-	-	0.85

Table 2: Summary of proposed metrics. 'HP' denotes headphones. 'L' denotes loudspeakers in a stereo setup. Numbers in the AB- and CD-ranges are in *Hz*. r represents the Pearson correlation coefficients. All numbers are rounded to two significant digits. Note that the correlations coefficients are not directly comparable as they are based on different validation schemes. * The AB and CD ranges (i.e. numerator and denominator respectively) and have been interchanged for this metric compared to the Paper. This was done to obtain a positive correlation. ** A lower correlation coefficient of $r = 0.70$ was found for the training data set.

The contributions of the literature study in Paper A consisted of descriptions of five focus areas of value for obtaining objective perceptual measurements, on which the prediction models could be trained. These were: methods of selecting the right sensory description and the proper number of these, as well as methods of loudness equalising sound reproduction systems, optimum listening room specifications for loudspeaker evaluation and finally an investigation of the validity of evaluations of loudspeakers using headphones in combination with auralization/virtualisation techniques.

In Paper B the main contribution was a study of the dominating percep-

tual characteristics differentiating headphones in general. Furthermore, the generic prediction model framework was established and tested on the two main MDS dimensions found.

In Paper C the framework from Paper B was again tested for headphones, but this time for prediction of perceptual characteristics defined by sensory descriptors from the Sound wheel [66]. Furthermore, the framework for prediction of the sensory descriptor Dark-Bright was proposed.

In Paper D both of the two frameworks were tested on predictions of perceptual characteristics of loudspeakers in a classical stereo setup. The generic framework was tested on four different sensory descriptors and the Dark-Bright framework was expanded by a proposed weighting functions, which lowered the influence of the midrange frequencies.

Prediction of rankings among systems
For this thesis (but not discussed in the papers), it was also investigated how well the metrics predicted the rankings of systems. Unfortunately, the relationship between the mean ratings of systems and the size of the confidence intervals meant, that the uncertainty of a systems ranking could be as high at six positions for 8 system$\times$sample combinations (loudspeaker validation set). A system could for example have a rank of 1 or 6 (*ranking uncertainty of 6*) depending on the true mean of the perceptual ratings of the set of loudspeakers. For the perceptual ratings of Paper C, the average ranking uncertainty was 1.8, and the number of correct ranking predictions within the uncertainty of the mean was 27/35 (77 %). The corresponding statistics for Paper D was an average ranking uncertainty of 4.5 and 24/24 (100 %) correct predictions. The worst prediction is terms of ranking was for Clean, were 4 of 7 predicted rankings were outside the range of possible rankings within the uncertainties of the perceptual data.

5.2 Secondary contributions

A number of secondary contributions were part of this thesis work as well. Presented below is a list of contributions selected on the basis of the 17 definitions of originality from [67, pp. 69-70].

- Introduced the notion that a prediction model might be needed in several versions to obtain accurate prediction for all types of sound reproduction systems (Section 1.1).
- Formulated a definition of listeners "internal reference", when evaluating perceptual characteristics (Paper D).
- Proposed a method for investigating uncertainties of small sets of sound reproduction systems using bootstrapping and categorisation (Paper C).

- Introduced, to the audio field of research, a set of criteria for gauging the suitability of perceptual rating data for predictive modelling, as well as more nuanced view of gauging model performance (Paper D).
- Introduced a method of monitoring the stability of pair-wise comparison data, such that data analysis can be performed on data from a sensible number of participants, which one can show is sufficient (Paper B).
- Discussed the potential value of prediction models with genre-dependent scaling in terms of perceived sensory descriptor intensities (Section 4.3).

6 Future work

A few points of interest that were not investigated within the allotted time of this project, were identified as having thepotential of leading to better metrics. First of all, the loudspeaker studies were conducted in a room with a intentionally limited reverberation time. In Paper A it was, however, shown that the reverberation time of common living rooms in western countries are longer than in DELTA's listening room, which follows the ITU-R BS.1116 standard [28]. The consequence of this mismatch, is a direct-to-reverberant (D/R) ratio, which is not generally representative. This have the effect of reducing the influence of directivity characteristcsof the loudspeakers, which have been found of high importance for preference ratings in [81] and successive studies by Floyd Toole and may also be a significant factor in discriminating between loudspeakers. Changing environment of evaluation of loudspeakers, may thus led to 1) better overall discriminating, 2) a higher influence of spatial properties, and 3) more representative (ecologically valid) listening test data and as a result, more relevant prediction models.

Another identified issue is, that for the loudspeaker metrics, the physical representation of the sound reproduction consists of recordings made in the listening position, which is assumed to be unique and fixed. Listeners, however, vary in height and will move their head. Consequently, it may be more suited to make a number of recordings which are representative of the range of positions of listeners' ears. A first approximation of this was attempted in the study described in Appendix III, but was not evaluated due to issue in the listening test design. A second attempt was made for the study of loudspeakers in a stereo setup. Here, the measurement positions were chosen on the basis of a study of listener head movements [37] as well as estimates of variation in listeners ear canals' height above floor level and an asymmetrical layout of recording positions chosen:

1. Centre / Sweet spot (height: 110 cm)

2. Centre (15° azimuth, right)
3. Centre (−30° azimuth, left)
4. Centre (height: 115 cm)
5. Centre (height: 101 cm)
6. Forward (7.5 cm from centre)

In Paper D only the sweet position was investigated, and as a result it remains to be investigated whether some position averaging scheme could improve the performance of the prediction models.

A final missing point of interest is whether prediction models could be made more generally valid, by calculating the metrics on the basis of pink noise or e.g. the IEC 60268-1 test signal [26], Either supplementary or as replacement of the musical excerpts. Pink noise have a frequency spectrum with equal energy between all octave bands, straining a loudspeaker or headphone evenly, while the IEC test signal is a filtered version of pink noise, which represents an average of reproduction material such as music and speech. A predict model calculated with a more general test signal, would reduce the cap between current traditional loudspeaker measurements and current perceptual measurement methods, and may increase the confidence in the prediction capabilities of the proposed metrics. Whether this would be possible, relies on whether a listening test with a small feasible amount of musical excerpts would be representative of the overall audio reproduction characteristics for a given target set of sound reproduction systems.

7 Concluding remarks

The purpose of this PhD project was to investigate the potential of modelling perceptual characteristics of headphones and loudspeakers using current state-of-the-art methods of conducting listening tests combined with incorporation of the current knowledge about the human auditory processing. The results of this project have demonstrated that this was at least possible with timbre-related sensory descriptors. Since these were also found to represent the two most dominating dimensions differentiating headphones (and loudspeakers [46]), underline the importance of these findings. In terms of other groups of sensory descriptors from the Sound wheel [66], the data quality analysis led to identified problems with all of them. Either in terms of mean ratings, which were not significantly different between the majority of evaluated systems, or due to lack of consistency in listener ratings, such as poor discriminating and/or repeatability. Whether this was a consequence of all other characteristics than the timbre-related ones being small (in comparison) or whether it was caused by issues in the listening test methodology remains to be investigated.

The modelling efforts of this project led to 12 metrics with Pearson correlations ranging from $r = 0.70 - 0.99$. Among these, bass was found consistently as a characteristic of importance regardless of the type of sound reproduction system and the method of perceptual evaluation: It was one of two dominating characteristics differentiating headphones (DeepBass), found as one of five reliable descriptors for prediction of differentiating characteristics among prototype headphones (Bass strength), as also found important for differentiating between loudspeakers in a stereo setup (BassPunch).

The high performance of these metrics not only showed the suitability of the data basis of the prediction models, but also of the methodology of collecting the data (recording technique and perceptual audio evaluation technique), as well and the processing of the recordings using loudness models mimicking the fundamental bottom-up auditory processing steps of the complex and highly non-linear human auditory system. Furthermore, the methodology was found suited for modelling of both headphones and loudspeakers, thus indicating that the influence of loudspeaker-room interaction affects neither the suitability of the modelling scheme nor the recording-based data collection method.

A concern in this project was to which degree it would be possible to make absolute predictions of perceptual evaluations, which was known to be relative in nature. The mechanics of listeners rating strategies and rating tendencies were described along with the possible measured to counter listeners' urge to use the rating scale in a non-absolute fashion. The two studies using sensory descriptors showed that it was possible to make predictions of the absolute distance in intensities of characteristics between systems. This, however, required linear fitting of the prediction models on the basis of perceptual ratings from individual musical excerpts; Strengthening both the prediction power of the final metrics, but also the dependence between the metrics and the original content of the listening tests used for training of the models.

The initial literature study of this project, showed that all previous modelling efforts had relied on the assumption of *one-size-fits-all*: That one prediction model could describe, e.g. preference, of all loudspeakers or headphones. The starting point of this thesis was, that this might not be the case. The results of the headphone study in Paper B, indicated that the two dominating dimension differentiating headphone might be closely related to the same dimensions for monophonic reproduction of loudspeakers in a room [46], but that a third dimension, "Feel of space", was not found in the headphone study. Combined with the findings of the remaining studies of this thesis as well as other studies (e.g. [2, 5, 18, 39, 85]), it shows that the differentiating characteristics varies between types of sound reproduction systems. Since preference must depend on these dominating characteristics, the assumption of *one-size-fits-all* is not likely to hold. Whether the perception of characteristics, which are similar between two or more types of systems,

differs in subtle ways depending on the type of reproduction systems, was not proven in this thesis work; A result of different sensory descriptors being chosen for characterisation in each study.

The methodologies described in this thesis may work for more than just the metrics proposed to this date. They constitute a framework for modelling of perceptual characteristics, which may be altered or adjusted in order to model more of the sensory descriptors of the Sound wheel in its present or future form. The loudness model may be replaced with future state-of-the-art models. The listening test methodologies may be optimised, e.g. by changing the reverberation time of the listening room. The generic modelling method (introduced in Paper B), may be altered for finding an optimum frequency range of perceptually relevant ripples in the STEP or the temporal averaging may be changed to capture dynamical aspects of the audio reproduction. In combination with future expansions of the suited data sets for modelling, this framework constitutes a sound foundation for future endeavours within perceptual modelling of sound reproduction characteristics.

References

[1] S. Bech and N. Zacharov, *Perceptual audio evaluation: theory, method and application*. Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2006.

[2] J. Berg and F. Rumsey, "Verification and correlation of attributes used for describing the spatial quality of reproduced sound," in *Audio Engineering Society Conference: 19th International Conference: Surround Sound - Techniques, Technology, and Perception*, Jun. 2001. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10057

[3] P. B. Brockhoff, P. Schlich, and I. Skovgaard, "Taking individual scaling differences into account by analyzing profile data with the Mixed Assessor Model," *Food Qual. Prefer.*, vol. 39, pp. 156–166, Jan. 2015.

[4] P. M. Brockhoff, "Assessor modelling," *Food Qual. Prefer.*, vol. 9, no. 3, pp. 87 – 89, 1998, sensometric Workshop.

[5] S. Choisel and F. Wickelmaier, "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *J. Acoust. Soc. Am.*, vol. 121, no. 1, pp. 388–400, 2007.

[6] S. B. Chon and K.-M. Sung, "Sound Quality Assessment of Earphone: A Subjective Assessment Procedure and an Objective Prediction Model," in *Proc. Audio Engineering Society Conference: 38: Sound Quality Evaluation*. Piteå, Sweden: Audio Engineering Society, Jun. 2010, pp. 1–8, paper No. 8-4.

[7] R. Conetta, F. Rumsey, P. Jackson, M. Dewhirst, S. Bech, D. Meares, and S. George, "QESTRAL (Part 2): Calibrating the QESTRAL Model Using Listening Test Data," in *Proc. of the Audio Engineering Society Convention 125*. San Francisco, CA, USA: Audio Engineering Society, 2008, pp. 1–18, convention paper 7596.

[8] C. Dehlholm, "Descriptive sensory evaluations, Comparison and applicability of novel rapid methodologies," Ph.D. thesis, University of Copenhagen, Copenhagen, Denmark, 2012, iSBN 978-87-7611-592-0.

[9] DELTA, Ed., *Celebrating the 100 year anniversary: Danish Loudspeakers*, first edition ed. Kgs. Lyngby, Denmark: Danish Sound Innovation, Dec. 2015. [Online]. Available: https://issuu.com/danishsound/docs/loudspeaker_100_year_aniversary_hig

[10] D. R. Derryberry, *Basic data analysis for time series with R*. Hoboken, New Jersey: Wiley, 2014.

[11] M. Dewhirst, R. Conetta, F. Rumsey, P. Jackson, S. Zielinski, S. George, S. Bech, and D. Meares, "QESTRAL (Part 4): Test Signals, Combining Metrics, and the Prediction of Overall Spatial Quality," in *Proc. of the Audio Engineering Society Convention 125*. San Francisco, CA, USA: Audio Engineering Society, 2008, pp. 1–8, convention Paper 7598.

[12] S. Fenton and H. Lee, "Towards a perceptual model of "Punch" in musical signals," in *Proc. of the Audio Engineering Society Convention 139*. New York, NY, USA: Audio Engineering Society, Oct. 2015, pp. 1–10, convention paper 9381.

[13] J. Francombe, R. Mason, M. Dewhirst, and S. Bech, "Investigation of a Random Radio Sampling Method for Selecting Ecologically Valid Music Program Material," in *Proc. of the Audio Engineering Society Convention 136*. Berlin, Germany: Audio Engineering Society, Apr. 2014, pp. 1–10, convention paper 9029.

[14] C. Fritz, A. F. Blackwell, I. Cross, J. Woodhouse, and B. C. J. Moore, "Exploring violin sound quality: Investigating English timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 783–794, 2012.

[15] A. Gabrielsson, "Perceived sound quality of reproductions with different frequency responses and sound levels," *J. Acoust. Soc. Am.*, vol. 88, no. 3, p. 1359, 1990.

[16] ——, "Loudspeaker frequency response and perceived sound quality," *J. Acoust. Soc. Am.*, vol. 90, no. 2, p. 707, 1991.

[17] A. Gabrielsson, U. Rosenberg, and H. Sjögren, "Judgments and dimension analyses of perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am.*, vol. 55, no. 4, p. 854, 1974.

[18] A. Gabrielsson and H. Sjögren, "Perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 1019–1033, 1979.

[19] Genesis, "History and description of loudness models. Loudness Toolbox for Matlab," Genesis S.A., Aix en Provence, France, Toolbox documentation IB/RP/10003, Dec. 2009.

[20] B. R. Glasberg and B. C. J. Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *J. Audio Eng. Soc*, vol. 50, no. 5, pp. 331–342, 2002.

[21] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[22] D. Hammershøi, R. Ordoñez, and A. T. Christensen, "Dose Estimate for Personal Music Players Including Earphone Sensitivity and Characteristic," in *Proc. of the Audio Engineering Society Conference: 2016 AES International Conference on Headphone Technology*. Aalborg, Denmark: Audio Engineering Society, Aug. 2016, pp. 1–7.

[23] N. Harlander, R. Huber, and S. D. Ewert, "Sound Quality Assessment Using Auditory Models," *J. Audio Eng. Soc*, vol. 62, no. 5, pp. 324–336, 2014.

[24] A. S. Harma, T. Lokki, and V. Pulkki, "Drawing Quality Maps of the Sweet Spot and Its Surroundings in Multichannel Reproduction and Coding," in *Proc. of the Audio Engineering Society Conference: 21st International Conference: Architectural Acoustics and Sound Reinforcement*. St. Petersburg, Russia: Audio Engineering Society, Jun. 2002, pp. 1–9, paper No. 64.

[25] J. Hots, J. Rennies, and J. L. Verhey, "Modeling Temporal Integration of Loudness," *Acta Acust. united Ac.*, vol. 100, no. 1, pp. 184–187, Jan. 2014.

[26] IEC, "Sound system equipment. Part 1. General," International Electrotechnical Commission (IEC), Geneva, Switzerland, Recommendation IEC 60268-1, 1985, edition 2.0.

[27] ITU-R, "Method for objective measurements of perceived audio quality," International Telecommunication Union Radiocommunication Assembly (ITU-R), United States, Recommendation ITU-R BS.1387-1, 2001.

[28] ——, "Recommendation BS 1116-3, Methods for the Subjective Assessment of Small Impairments in audio Systems Including Multichannel Sound Systems." International Telecommunication Union Radiocommunication Assembly (ITU-R), United States, Recommendation ITU-R BS 1116-3, Feb. 2015.

[29] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU Telecommunication Standardization Sector (ITU-T), United States, Recommendation ITU-T PP.862, 1999.

[30] ITU-T, "Perceptual objective listening quality assessment," ITU Telecommunication Standardization Sector (ITU-T), United States, Recommendation ITU-T P.863, Jan. 2011.

[31] P. Jackson, M. Dewhirst, R. Conetta, and S. Zielinski, "Estimates of Perceived Spatial Quality across the Listening Area," in *Proc. of the Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*. Piteå, Sweden: Audio Engineering Society, Jun. 2010, pp. 1–10, paper No. 8-1.

[32] P. Jackson, M. Dewhirst, R. Conetta, S. Zielinski, F. Rumsey, D. Meares, S. Bech, and S. George, "QESTRAL (Part 3): System and Metrics for Spatial Quality Prediction," in *Proc. of the Audio Engineering Society Convention 125*. San Francisco, CA, USA: Audio Engineering Society, Oct. 2008, pp. 1–9, convention paper 7597.

[33] M. L. Jepsen and T. Dau, "Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss," *J. Acoust. Soc. Am.*, vol. 129, no. 1, pp. 262–281, 2011.

[34] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 422–438, 2008.

[35] D. H. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J. Acoust. Soc. Am.*, vol. 68, no. 4, pp. 1115–1122, 1980.

[36] A. J. M. Kaizer and A. Leeuwestein, "Calculation of the Sound Radiation of a Nonrigid Loudspeaker Diaphragm Using the Finite-Element Method," *J. Audio Eng. Soc*, vol. 36, no. 7/8, pp. 539–551, 1988.

[37] C. Kim, R. Mason, and T. Brookes, "Head Movements Made by Listeners in Experimental and Real-Life Listening Activities," *J. Audio Eng. Soc*, vol. 61, no. 6, pp. 425–438, 2013. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16833

[38] S. Kim and W. L. Martens, "Deriving Physical Predictors for Auditory Attribute Ratings Made in Response to Multichannel Music Reproductions," in *Proc. of the Audio Engineering Society Convention 123*. New York, NY, USA: Audio Engineering Society, Oct. 2007, pp. 1–10, convention paper 7195.

[39] W. Klippel, "Multidimensional Relationship between Subjective Listening Impression and Objective Loudspeaker Parameters," *Acta Acust. united Ac.*, vol. 70, no. 1, pp. 45–54, 1990.

[40] J. Kruskal and M. Wish, *Multidimensional scaling*. California, USA: Sage Publications, 1978.

[41] M. Kumagai, Y. Suzuki, and T. Sone, "A study on the time constant for an impulse sound level meter," *J. Acoust. Soc. JPN*, vol. 5, no. 1, pp. 31–36, 1984.

[42] A. Kuusinen, J. Pätynen, S. Tervo, and T. Lokki, "Relationships between preference ratings, sensory profiles, and acoustical measurements in concert halls," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 239–250, Jan. 2014.

[43] I. B. Labuschagne and J. J. Hanekom, "Preparation of stimuli for timbre perception studies," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2256–2267, 2013.

[44] M. Lavandier, P. Herzog, and S. Meunier, "Comparative measurements of loudspeakers in a listening situation," *J. Acoust. Soc. Am.*, vol. 123, no. 1, pp. 77–87, 2008.

[45] M. Lavandier, S. Meunier, and P. Herzog, "Perceptual and physical evaluation of differences among a large panel of loudspeakers," in *Forum Acusticum 2005 Budapest proceedings*. Budapest, Hungary: Acta Acustica United with Acustica, Aug. 2005, pp. S111–2.

[46] ——, "Identification of some perceptual dimensions underlying loudspeaker dissimilarities," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4186–4198, 2008.

[47] W. M. Leach, *Introduction to electroacoustics & audio amplifier design*, 4th ed. Dubuque, IA: Kendall/Hunt Publishing Company, 2010, oCLC: 703226032.

[48] G. Lorho, "Subjective Evaluation of Headphone Target Frequency Responses," in *Proc. of the Audio Engineering Society Convention 126*. Munich, Germany: Audio Engineering Society, May 2009, pp. 1–20, convention paper 7770.

[49] G. Lorho, G. Le Ray, and N. Zacharov, "eGauge—A Measure of Assessor Expertise in Audio Quality Evaluations," in *Proc. of the Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*. Piteå, Sweden: Audio Engineering Society, Jun. 2010, pp. 1–10, paper No. 7-2.

[50] E. A. Macpherson, "A Computer Model of Binaural Localization for Stereo Imaging Measurement," *J. Audio Eng. Soc*, vol. 39, no. 9, pp. 604–622, 1991.

[51] R. Mason and F. J. Rumsey, "A Comparison of Objective Measurements for Predicting Selected Subjective Spatial Attributes," in *Proc. of the Audio Engineering Society Convention 112*. Munich, Germany: Audio Engineering Society, Apr. 2002, pp. 1–18, convention paper 5591.

[52] P.-Y. Michaud, M. Lavandier, S. Meunier, and P. Herzog, "Objective Characterization of Perceptual Dimensions Underlying the Sound Reproduction of 37 Single Loudspeakers in a Room," *Acta Acust. united Ac.*, vol. 101, no. 3, pp. 603–615, May 2015.

[53] B. C. J. Moore, "Development and Current Status of the "Cambridge" Loudness Models," *Trends in Hearing*, vol. 18, pp. 1–29, Oct. 2014.

[54] B. C. J. Moore and B. R. Glasberg, "Modeling binaural loudness," *J. Acoust. Soc. Am.*, vol. 121, no. 3, pp. 1604–1612, 2007.

[55] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *J. Audio Eng. Soc*, vol. 45, no. 4, pp. 224–240, 1997.

[56] B. C. J. Moore and C.-T. Tan, "Development and Validation of a Method for Predicting the Perceived Naturalness of Sounds Subjected to Spectral Distortion," *J. Audio Eng. Soc*, vol. 52, no. 9, pp. 900–914, 2004.

[57] B. C. J. Moore, C.-T. Tan, N. Zacharov, and V.-V. Mattila, "Measuring and Predicting the Perceived Quality of Music and Speech Subjected to Combined Linear and Nonlinear Distortion," *J. Audio Eng. Soc*, vol. 52, no. 12, pp. 1228–1244, 2004.

[58] B. C. Moore, "Temporal integration and context effects in hearing," *J. Phonetics*, vol. 31, no. 3-4, pp. 563–574, Jul. 2003.

[59] D. S. Moore, G. P. McCabe, and B. A. Craig, *Introduction to the practice of statistics*, 6th ed. New York: W.H. Freeman, 2014.

[60] S. Olive and T. Welti, "The Relationship between Perception and Measurement of Headphone Sound Quality," in *Proc. of the Audio Engineering Society Convention 133*. San Francisco, CA, USA: Audio Engineering Society, Oct. 2012, pp. 1–17, convention Paper 8744.

[61] S. E. Olive, "A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model," in *Proc. of the Audio Engineering Convention 117*. San Francisco, CA, USA: Audio Engineering Society, 2004, pp. 1–21, convention paper 6190.

[62] S. L. Olsen, F. T. Agerkvist, E. MacDonald, T. Stegenborg-Andersen, and C. P. Volk, "Modeling the Perceptual Components of Loudspeaker Distortion," in *Proc. of the Audio Engineering Society Convention 140*. Paris, France: Audio Engineering Society, Jun. 2016, pp. 1–9, convention paper 9549.

[63] M. Opitz, "Headphones Listening Tests," in *Proc. of the Audio Engineering Society Convention 121*. San Francisco, CA, USA: Audio Engineering Society, Oct. 2006, pp. 1–12, convention paper 6890.

[64] T. H. Pedersen and C. Fog, "Optimisation of perceived product quality," in *Euronoise 98 Designing for Silence*, vol. II. Hannover: UB/TIB Hannover, 1998, pp. 633–638.

[65] T. H. Pedersen and N. Zacharov, "How many psycho-acoustic attributes are needed?" in *Acoustics'08 Paris: June 29 - July 4, 2008.* Paris, France: Société Francaise d'Acoustique (SFA), 2008, pp. 1215–1220.

[66] ——, "The Development of a Sound Wheel for Reproduced Sound," in *Audio Engineering Society Convention 138*. Warsaw, Poland: Audio Engineering Society, May 2015, pp. 1–13, convention Paper 9310.

[67] E. M. Phillips and D. S. Pugh, *How to get a PhD: A handbook for students and their supervisors*, 5th ed. Maidenhaed, Berkshire, England: Open University Press, Oct. 2010.

[68] T. Poulsen, "Loudness of tone pulses in a free field," *J. Acoust. Soc. Am.*, vol. 69, no. 6, pp. 1786–1790, 1981.

[69] E. C. Poulton, *Bias in quantifying judgements*. Hove u.a: Lawrence Erlbaum, 1989.

[70] F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, R. Conetta, S. George, S. Bech, and D. Meares, "QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction Using an Artificial Listener," in *Proc. of Audio Engineering Society Convention 125*. San Francisco, CA, USA: Audio Engineering Society, 2008, pp. 1–8, convention Paper 7595.

[71] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences," *J. Acoust. Soc. Am.*, vol. 117, no. 6, pp. 3832–3840, 2005.

[72] E. Schubert and J. Wolfe, "Does Timbral Brightness Scale with Frequency and Spectral Centroid?" *Acta Acust. united Ac.*, vol. 92, pp. 820–825, 2006.

[73] A. Silzle, B. Neugebauer, S. George, and J. Plogsties, "Binaural Processing Algorithms: Importance of Clustering Analysis for Preference Tests," in *Proc. of the Audio Engineering Society Convention 126*. Munich, Germany: Audio Engineering Society, May 2009, pp. 1–17, convention paper 7728.

[74] R. H. Small, "Closed-Box Loudspeaker Systems-Part 1: Analysis," *J. Audio Eng. Soc*, vol. 20, no. 10, pp. 798–808, 1972.

[75] ——, "Closed-Box Loudspeaker Systems-Part 2: Synthesis," *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 11–18, 1973.

[76] G. A. Soulodre, "Evaluation of Objective Loudness Meters," in *Proc. of the Audio Engineering Society Convention 116*. Berlin, Germany: Audio Engineering Society, May 2004, pp. 1–12, convention paper 6161.

[77] H. Staffeldt, "Measurement and prediction of the timbre of sound reproduction," *J. Audio Eng. Soc*, vol. 32, no. 6, pp. 410–414, Jun. 1984.

[78] M. Takanen and G. Lorho, "A Binaural Auditory Model for the Evaluation of Reproduced Stereophonic Sound," in *Proc. of the Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*. Helsinki, Finland: Audio Engineering Society, Mar. 2012, pp. 1–10.

[79] C.-T. Tan, B. C. J. Moore, and N. Zacharov, "The Effect of Nonlinear Distortion on the Perceived Quality of Music and Speech Signals," *J. Audio Eng. Soc*, vol. 51, no. 11, pp. 1012–1031, 2003.

[80] C.-T. Tan, B. C. J. Moore, N. Zacharov, and V.-V. Mattila, "Predicting the Perceived Quality of Nonlinearly Distorted Music and Speech Signals," *J. Audio Eng. Soc*, vol. 52, no. 7/8, pp. 699–711, 2004.

[81] F. E. Toole, "Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2," *J. Audio Eng. Soc*, vol. 34, no. 5, pp. 323–348, 1986.

[82] T. Worch, S. Lê, and P. Punter, "How reliable are the consumers? Comparison of sensory profiles from consumers and experts," *J. Food. Qual. Prefer.*, vol. 21, no. 3, pp. 309–318, Apr. 2010.

[83] T. Worch, S. Lê, P. Punter, and J. Pagès, "Ideal Profile Method (IPM): The ins and outs," *Food Qual. Prefer.*, vol. 28, no. 1, pp. 45–59, Apr. 2013.

[84] N. Zacharov, T. Pedersen, and C. Pike, "A common lexicon for spatial sound quality assessment - latest developments," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. Lisbon, Portugal: IEEE, Jun. 2016, pp. 1–6.

[85] N. Zacharov, J. Ramsgaard, G. Le Ray, and C. V. Jørgensen, "The multidimensional characterization of active noise cancelation headphone perception," in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*. Trondheim, Norway: IEEE, Jun. 2010, pp. 130–135.

[86] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review," *J. Audio Eng. Soc*, vol. 56, no. 6, pp. 427–451, 2008.

Part II

Papers

Paper A

Five aspects of maximizing objectivity from perceptual evaluations of loudspeakers: A literature study

Christer P. Volk, Søren Bech,
Torben H. Pedersen, and Flemming Christensen

The paper has been published in the
Proc. of Audio Engineering Society Convention 138, pp. 1—12.
Warsaw, Poland: Audio Engineering Society, 2015.

To accommodate the smaller format of the printed thesis publication, the layout has been revised. The text is identical to the submitted/published article.

Abstract

A literature study was conducted focusing on maximizing objectivity of results from listening evaluations aimed at establishing the relationship between physical and perceptual measurements of loudspeakers. The purpose of the study was to identify and examine factors influencing the objectivity of data from the listening evaluations. This paper addresses the following subset of aspects for increasing the objectivity of data from listening tests: The choice of perceptual attributes, relevance of perceptual attributes, choice of loudness equalisation strategy, optimum listening room specifications, as well as loudspeaker listening in-situ vs. listening to recordings of loudspeakers over headphones.

1 Introduction

In the natural sciences, a fundamental prerequisite for making valid conclusions is the collection of objective results from measurements. With regards to perceptual evaluations, using human subjects, the same goal applies. In [1] by Blauert, the requirement of objectivity is defined as results which are always the same, with regards to the statistical inferences, both for multiple measurements of one assessor and measurements of multiple assessors. Two factors constitutes this requirement: reproducibility and accuracy. In this context, accuracy is defined as closeness to the *true* answer and reproducibility is defined as the degree of variability around this *true* answer.

Maximizing the objectivity of results in listening evaluations (of loudspeakers) is a matter of controlling the experimental variables, both the physical and the psychological. For an overview of these see e.g. [2, 3] or one of the many listening test standards. Among the most commonly mentioned are: 1) Playback systems, 2) Attributes, 3) Listening test paradigm, 4) Test management, 5) Listening room, 6) Stimuli, 7) Listening panel, and finally 8) Statistics.

The purpose of the present study was to ensure optimal data for an industrial Ph.D. project aimed at modelling the statistical relationship between electro-acoustical measurements and perceptual evaluations in the domains of loudspeakers, headphones, and portable audio systems. Through a comprehensive literature study, looking into the aforementioned experimental variables in more detail, a number of aspects were identified as necessary for maximizing objectivity of perceptual evaluation. Five of these aspects are described in the present paper.

Among the physical variables three are investigated in this study: Listening room specifications, Loudness equalisation, as well as listening mode, i.e. listening in-situ to loudspeakers versus listening to loudspeaker recordings over headphones. The first two have been found to have a statistically

significant influence on the results of listening tests (see e.g. [4–7]). The last variable, listening mode, will in case of reproduction over headphones, have a significant impact on the results if the headphones has not been adequately calibrated/equalized.

Beside the physical variables a number of dependent variables in the perceptual measurements are also of importance with regards to objectivity, i.e. variables related to the experimental questions of interest (and the method of enquiring). The information needed from perceptual evaluation of loudspeakers typically relates to either: Knowledge about the basic audio quality (which loudspeaker has the highest audio quality?), the perceptual differences (which acoustical differences are audible?), or the perceptual differences significantly influencing basic audio quality (which perceptual differences are important for preference?). In the third case, the group of sensory descriptors (timbre, spatial, dynamics, etc.) and the selection of attributes within these groups has an influence of the prediction of preference, and therefore becomes directly related to the level of accuracy (and thus the objectivity) of the results.

Notice however, that other factors, such as defining sensory descriptors as well as training and monitoring of assessors are also major issues, related to reduction of statistical noise in the evaluations, but are outside the scope of this paper. Interested readers are encouraged to read the overviews of sensory descriptors elicitation methodologies in the Ph.D. thesis' of Lorho [8] and Dehlholm [9].

The next sections present studies of each of the five aspects, and finally a section with a discussion of conclusions.

2 Sensory descriptors

Within sensory evaluations, it is not yet clear how many descriptors are needed to measure all (significant) perceptual characteristics within the range of reproduced sound from loudspeakers. In this context a perceived characteristic (of e.g. an acoustical source) is considered evidence of an perceptual attribute. In many cases it is possible to label a perceptual attribute with a word, the sensory descriptor. A large number of these sensory descriptors are suggested in the literature (scientific and commercial) to describe sensations of loudspeakers - both perceptual and affective. A low number of attributes is preferable with regards to simplicity, but choosing a set of sensory descriptors, which does not relate to all perceivable differences between the products being evaluated, leads to a risk of bias, which reduces the accuracy of the evaluation. This bias stems from the attributes not included, which may influence the ratings of the perceptual attributes included (see e.g. [3]). An example could be perceived distortion affecting an assessor's

rating of 'clarity'. The phenomenon is sometimes referred to as perceptual 'bleeding'.

The sets of sensory descriptors reported in the literature for audio reproduction systems, such as loudspeakers and headphones ranges in number from 7 sensory descriptors in one study up to 55 in another, which could therefore be considered the range of expected sensory descriptors required to comprehensively describe the perceptual characteristics of a set of loudspeaker or a set of headphones [10–13].

Basic audio quality (BAQ) is assumed to be the weighted auditory summation of a stimulus' contributions from individual perceptual attributes [3]. In an effort to predict the perceptual attributes involved in the summation, one approach is to make a statistical regression model of the measured responses of the evaluated sensory descriptors, and investigate the relationship to the affective metric BAQ. A regression model can provide insight with regards to the BAQ variance explained by the individual sensory descriptors' ratings, as illustrated in Fig. A.1. If the perceptual attributes described by the sensory descriptors are independent and unidimensional, the number of principal components/dimensions in the data should correspond to the number of sensory descriptors. If these requirements are not fully met, the number of principal dimensions will be lower than the number of descriptors. In one study by Rumsey et al. [14] concerning multi-channel reproduction and utilizing the regression model approach, 2 main dimensions was found to explain 97 % of the variance in an experiment with BAQ prediction of sound reproduced by loudspeakers.

Another approach is useful if the perceivable differences between loudspeakers' sound reproduction capabilities are of interest (independent of whether they contribute significantly to perception of BAQ). An evaluation is conducted measuring the differences between loudspeakers (e.g. a pair-wise comparison test or an attribute test) and a statistical investigation (such as PCA or MDS) is made of the dimensional span of the evaluations. Among the experiments described in [10, 12, 15, 16] ([12] was a headphone experiment and [10] included both a loudspeaker and a headphone experiment), 2-4 main dimensions explained most of the variance in experiments of dissimilarities of sound reproduced by loudspeakers:

- 2 dimensions in [15] (explained variance not reported)
- 87 % for 2 dimensions in [16]
- 90 % for 4 dimensions in [10]

For the two cited papers on headphones the results were similar with 85 % reported for 5 dimensions in [10], and 96 % for 4 dimensions in [12].

Within the methodologies described so far lies the assumption that perceptual attributes can be adequately described by using only words as sen-

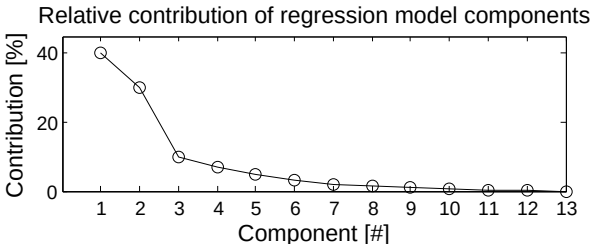


Fig. A.1: Example of variance explained by each component in a regression model. In this example six components are needed in the model to explain 95 % of the variance in the ratings of a basic audio quality evaluation.

sory descriptors. This assumption has previously been studied with the hypothesis that a graphical response format might lead to a more accurate description of especially spatial perceptual attributes. This was for instance investigated in [17, 18], where a graphical response format was found to be a useful supplement to the verbal attributes, i.e. leading to a more detailed spatial characterisation, fewer attributes, and was found to reveal collinearity problems within the verbal sensory descriptors related to spatial perception[1]. The conclusion being that the limitations of words as mediators of spatial perception, are important to take into account in evaluations where the spatial perception is of interest.

Summing up, the number of sensory descriptors needed to describe perceptual loudspeaker differences may be as low as 2-4 if carefully selected, i.e. 1) having a high degree of independence (and unidimensionality) among them, 2) sufficiently describing the perceptible loudspeaker differences, and 3) being sufficient to describe the perceptible differences of interest (as opposed to including a graphical response format instead of words).

3 Relevant perceptual attributes

A common trait regarding known methods of elicitation of sensory descriptors for characterisation of loudspeaker differences (such as QDA [19]), is the emphasis on uncovering all perceived differences, rather than on identifying the dominating perceivable differences and their individual importance. Furthermore the result is often presented in a spider plot, where the importances of each descriptor with regards to basic audio quality (or preference) is missing. One exception is the Napping method [20], where the limitation of two dimensions (a piece of paper) and the difference in width and height forces

[1]Collinearity in statistics is the occurrence of linear relations between variables leading to problems in the calculation of regression coefficients.

the assessor to actively choose the most importance ones. For loudspeaker evaluations the limitation of only two dimensions may however be too restrictive (in accordance with the conclusions of section 2). In the next subsection studies investigation the salience of perceptual attributes are discussed.

3.1 Salience of perceptual attributes

One approach to uncover which perceptual attributes dominates the overall perception of basic audio quality is to investigate the auditory sensitivity within the areas of acoustics wherein loudspeakers differentiate: changes in spectrum, phase, source directivity etc. The outcome of this approach is briefly reviewed by Brian Moore in [21] (section 10.3). His main points are summarized in the next few lines. The lack of spectral distortion (here defined as all deviations from a flat frequency response) has been found to correlate highly with preference. Not only regarding the on-axis frequency response, but also off-axis (in the horizontal and median plane), which contributes with lateral energy and thus the sense of a spatial reproduction. Besides the sensitivity to spectral distortion, Moore touches on the subject of phase distortion (deviations from perfect phase reproduction). The human auditory system is very sensitive to phase distortion, especially with regards to transients, as the stimuli reaching the ear is processed/perceived without the presence of reflections, as these reaches the ear with a time delay larger than the duration of the transients.

Other researchers have looked into sensitivity in the perception of sound quality of loudspeakers as well: The perceptual influence of frequency response irregularities were investigated in 1981 by Bücklein [22], phase distortion was investigated by Møller in [23] and again in [24], and by Choisel and Martin for headphones [25] and for loudspeakers [26], while Bech investigated audibility of low-frequency irregularities in [27]. Additionally, an auditory model was described in a number of papers from 2003-2004 by Tan et al., which was trained to predict perceptual degradation of common spectral distortions in loudspeakers [28, 29] (validated in [30]).

These studies of the audibility of degradations in loudspeaker reproduction, using a psychoacoustics approach may well prove useful for understanding the salience of perceptual attributes.

3.2 The role of timbre

While the quantity and importance of sensory descriptors depend on the type of loudspeakers and loudspeaker setup (for example the number of reproduction channels), timbre has been identified as the dominating factor in a number of studies, ranging from an experiment using a mono loudspeaker setup investigating perceived differences in loudspeaker comparisons [15],

to a surround sound experiment investigating contributions from perceptual attribute categories to the overall perception of basic audio quality [14]. In the study by Rumsey et al. [14] it was established that approximately 70 % of the perceptual differences between loudspeakers were related to timbre for surround sound stimuli (5.0) down-mixed to various configurations of three, two or one loudspeaker.

The importance of timbre is not limited to loudspeakers, but was for instance also found to be dominating in a headphone study by Olive et. al. [13], where the sensory descriptor "Good spectral balance" was found to have the highest correlation with Preference ($r = 0.92$).

To maximize the objectivity of a study accessing the general characteristics or main differences of loudspeakers' sound reproduction, it is inferred from the mentioned studies, that the set of sensory descriptors must be chosen such that it is dominated by descriptors related to timbre.

4 Loudness adjustment strategy

Differences in loudness between the sound reproduction of different loudspeakers has long been known to be a confounding factor in perceptual evaluation of sound quality [6, 7], i.e. affecting evaluations of other attributes if not properly controlled. Therefore loudness equalisation is needed to avoid a systematic bias in perceptual evaluations. In principle the ideal loudness equalisation should account for differences in loudspeakers' sound reproduction (sensitivity as well as frequency- and phase response), samples/programme material, and assessors' loudness perception and any interaction among these.

An inherent problem with loudness equalisation, is stimuli containing multiple sources, such as is common for music. While the overall perception of a particular equalisation scheme might lead to a set of stimuli being perceived as equally loud on a set of loudness equalised loudspeakers, the individual sources in the stimuli might not, e.g. the bass may seem louder on some of the loudspeaker and the guitar lower on others. This is caused by differences in frequency responses, i.e. dips and peaks. While the dips and peaks may differ in frequency and magnitude from one loudspeaker to another, all of them have dips at the very low and very high frequencies - the roll-off. This fact makes it relevant to decide whether the loudness equalisation scheme should include the entire audible frequency range, or only the frequency range within the lowest common cut-off frequency of the set of loudspeakers.

Furthermore large differences in loudness perception exists between subjects, e.g. as a consequence of natural variation in hearing thresholds [31]. This nature variation is allowed in perceptual evaluations, as it is common

in perceptual evaluations to allow assessors within the entire normal hearing range up to $\leq 20\,dB\,HL$ [32] or even $\leq 25\,dB\,HL$ [33]. While assessors fulfilling these lax requirements may be representative of the general population, it allows for assessor differences not only with regards to absolute differences in overall loudness perception, but also with regards to frequency dependent differences. The difference in auditory threshold becomes especially relevant, when the stimuli included in the listening evaluation have a dynamic range with parts in the range close to the hearing threshold. Combined with the fact that differences in loudness was found to be perceived for small differences in sound levels (Just noticeable differences of $\approx 0.5\ dB\,SPL$) [34], sufficiently good loudness equalisation becomes a challenge to accomplish.

Over the years many loudness equalisation schemes have been utilized to deal with the challenges of loudness equalisation. These can be categorised as listed here:

1. Physical RMS measurements with various time- and frequency weightings (e.g. using a pink noise calibration signal as in [35])

2. Physical measurements of loudness based on various auditory loudness models (e.g. models by either Zwicker [36, 37] or Moore [30, 38])

3. Perceptual measures (e.g. averaged loudness equalisation from perceptual listening tests as in [25])

For each of these three methods the calibration signal can be chosen as a general signal such as Pink noise (broadband or narrowband) or the set of stimuli used in the main listening test.

The approach described in item 1 was recently used at DELTA (method described in [39]). A band-limited pink noise reference stimulus was played back on the loudspeakers being tested and the equivalent sound pressure levels, L_{eq}, were measured and adjusted to the same level. The result was an experimental setup with low-frequency content having audible differences in loudness for selected musical excerpts, but having approximately the same perceived loudness of the vocal.

The choice of frequency weighting scheme (A, B, ...) used for loudness equalisation using a pink noise test signal has been discussed in the literature as well. In [35] it is reported that equalising using a B-weighted pink noise test signal correlates better with perception than A-weighted pink noise test signals in tasks of loudness equalisation performed by individual assessors.

A newer paper by Soulodre and Norcross [34] recommends another weighting curve, 'RLB', which is a cross between the standard 'B' and 'C' curves at low frequencies and flat above $\approx 400\,Hz$, which is found to better match subjective loudness ratings of 'typical program material', compared to the standard A- and B- weighting curves.

Using loudness models to equalise loudness across loudspeakers is not presently found perceptually satisfactorily for complex stimuli such as music, but some papers describe a procedure, where the models are used to obtain a rough approximation followed by a manual 'fine-equalization' of the loudspeakers' sound reproduction by the experimenters or a subset of the assessors (e.g. [14]).

Recently, Koehl and Paquier [40] (and later Koehl et al. [41]) have suggested a novel method, which attempts to eliminate individual perceptual differences of loudness from listening tests. By allowing the assessors to adjust the playback level individually within $\pm$ 6 dB with the instruction of equalising the playback level of each stimulus. This strategy was compared to a strategy of loudness equalising by having three expert listeners agree on loudspeaker levels for the individual music excerpts across the four loudspeakers. Their results showed comparable results in a preference evaluation, but that the assessors were significantly better at discriminating between some of the loudspeakers, in the sessions where the individual assessors equalised the loudspeakers.

While it may seem intuitively better to loudness equalise loudspeakers for each stimulus in a loudness matching experiment, Soulodre et al. showed in [34] that performance in terms of assessor reliability had the smallest standard error for white noise stimuli compared to a variety of program material (jazz, speech, etc.).

Furthermore, for narrow band stimuli within frequency ranges coinciding with significant loudspeaker differences, the loudspeaker differences may be unintentionally compensated for, basically comprising a frequency response equalisation rather than a loudness equalisation.

5 Listening room specifications

Since the British Broadcasting Corporation (BBC) started investigating the influence of domestic room acoustics on the perceived quality of their transmissions [42], many approaches have been made to obtain results of listening tests, which could be considered general and valid for the largest possible proportion of domestic listening spaces.

The room influence is a general concern in relation to perceptual evaluation of loudspeakers, as the loudspeaker-room-listener interaction is evaluated rather than the loudspeaker isolated. Therefore the evaluation setup should mimic that of the intended users of the loudspeakers being evaluated. A loudspeaker will be used in many acoustically different rooms, positions, and conditions, all having an influence on the sound field. Furthermore the users cannot be assumed to listen solely in the sweet spot[2]. They will sit in a

[2]The sweet spot is the optimum listening position, which the sound experience is commonly

large range of positions relative to the loudspeakers; positions varying both horizontally and vertically, and often the listeners will not even be stationary, but moving around. How do we evaluate a loudspeaker such that the conclusions are valid for the largest possible proportion of listening spaces? The current standards (described later on) strives to achieve a room response which has limited colouration (changes in spectral composition of a played back stimulus compared to the original) in the mid- and high frequencies and a room which is symmetrical in an effort to obtain an unbiased evaluation. The underlying assumption being that this strategy will correlate the most with the average listening experiences of the users.

Two main listening standards are commonly utilized in perceptual loudspeaker evaluations: The ITU standard ITU-R BS.1116-1 [43] and the IEC 60268-13 standard [44]. Other standards include the AES20 standard [45] which has specifications similar to the IEC standard, and the EBU Tech 3276 [46, 47] which closely resembles the ITU standard with regards to room specifications. ITU-R BS.1116-1 [43] was designed for critical listening and specifically for evaluation of codecs and detection of small degradations. It specifies a very low reverberation time, which minimises the room influence and put emphasis on the direct sound. A concern with listening rooms fulfilling this strict standard is, that few domestic rooms have reverberation times this low, which has the consequence that loudspeakers designed for domestic conditions will sound different in most real life situations. A better alternative might be the IEC 60268-13 standard [44], which has the same philosophy regarding colouration, but specifies a larger target for the reverberation times. The reverberation times in the IEC standards corresponds to those found in the mentioned BBC study [42]. Utilizing the more reverberant IEC 60268-13 standardised listening room, will affect the spatial sensation of the stimuli and increase the perceptual difference between loudspeakers with differences in directivity patterns, as well as being assumed to correspond better to a common western living room.

Studies of domestic rooms in Europe (UK: [42, 48, 49], Spain: [50], Austria: [51]) can provide background knowledge for making a qualified choice between the room specifications from ITU and IEC. A list of prior studies of reverberation times and room sizes is presented in Table A.1, dating from the aforementioned BBC study from the 1950s [42] to a recent and very comprehensive survey by Dias & Pedrero from 2005 [50].

From Table A.1 the difference in reverberation times between the Building Research Establishment (BRE) study in UK and the Diaz & Pedrero study [50] in Spain showed the influence of very different styles of buildings and interior decoration. Notice that even the low reverberation times of the British

optimised for, e.g. for a stereo setup the third corner of a equal-sided triangle with loudspeakers in the other two corners.

Author & source	Year	Rooms [#]	Avg. volume [m^3]	Reverberation time [s]
Geddes et al. [42]	1954	16	Not reported	$60/8000\,Hz : 0.55/0.30$
Jackson & Leventhall [48]	1972	50	44	$125/8000\,Hz : 0.69/0.40$
BRE [49]	1968	14	Not reported	$100/3150\,Hz : 0.39/0.32$
BRE [49]	1972	60	35	$100/3150\,Hz : 0.44/0.33$
BRE [49]	1983	47	39	$100/4000\,Hz : 0.38/0.37$
Diaz & Pedrero [50]	2005	3211	46.2	$125/4000\,Hz : 0.60/0.35$
Lang, Judith [51]	2012	113	30-80	$63/3150\,Hz : 0.60/0.48$
ITU-R BS.1116-1 [43]	1997		100*	$200\text{-}4000\,Hz : 0.25 \pm 0.05$
IEC 60268-13 [44]	1998		100.17**	$200\text{-}4000\,Hz : 0.45 \pm 0.15$

Table A.1: Surveys of reverberation times in domestic rooms (1954-2005). Maximum and minimum reverberation time with the specified frequency range is reported. The two bottom rows displays the two main listening room standards. * The ITU standard reverberation times are stated here for a $100\,m^2$ room, but other sizes are allowed. For smaller room sizes the allowed reverberation times are lower. ** The IEC reference room volume is stated here. In the IEC standard the specified reverberation times are contant for all allowable room sizes.

domestic rooms are higher than those specified in the ITU standard ITU-R BS.1116-1 [43].

While the ITU standard might be well suited for critical evaluation of codecs, the requirements of the IEC 60268-13 standard [44] was found to be a better fit to the reverberation times reported in Table A.1 of domestic rooms in Europe.

6 Listening in-situ vs. listening over headphones

When conducting direct perceptual evaluations of loudspeakers, they should ideally be evaluated under identical conditions. To ensure this, the loudspeakers should be positioned in the same place in the room or the influence of different positions should be eliminated. Furthermore the limited human auditory memory (limited in time to $\approx 20\,sec$ [52]), should be accounted for in the measurement setup, i.e. by enabling the assessors to switch between loudspeakers fast or ideally instantaneously. Consequently, the ideal is to have the loudspeakers positioned in the same position at the same time. As this is not possible a compromise is needed. One possibility is to have a loudspeaker shuffler to swiftly switch between loudspeakers, enabling an measurement setup with a fixed loudspeaker position - for the loudspeaker playing. The shuffler must operate silently to avoid unwanted cues. An alternative is to position the loudspeakers side-by-side and randomise their

position between assessors in order to statistically reduce the variation of room influence between positions, affecting the evaluation accuracy. In this setup, the loudspeakers do however influence each others acoustical output to some degree. Lastly, the room influence of different loudspeaker positions (with sufficient distances to minimize interaction) can be included in the design of the experiment, in an effort to statistically separate the influence of the loudspeaker-room interaction from the loudspeaker output. While this is still considered a method of direct evaluation, having loudspeaker position as an experimental variable increases the number of assessors (or listening sessions) required to ensure a statistically balanced experiment.

Indirect listening is an alternative to direct listening. In indirect listening evaluations recordings are made of each loudspeaker, e.g. using a head-and-torso simulator (HATS). Afterwards the recordings are reproduced over the assessors' headphones. This allows all loudspeakers to be evaluated while positioned ideally. Furthermore the assessor position is well-defined, as the position of the recording microphones does not change during playback, while an assessor in a direct comparison could move during a listening session.

It also allows multiple assessors to make evaluations of the same loudspeakers simultaneously, as the evaluation does not require use of the physical loudspeakers, which is of practical value. But is it a valid method for evaluation of loudspeakers? Does it lead to the same conclusions as the direct evaluations? And under which circumstances? The literature review for the considerations listed below included the following topics: Headphone transfer functions [53], Headphone calibration ([3], section 8.3), Binaural Synthesis [11, 54–56], Assessor asymmetry [57], Localisation performance [58], and Auralization of loudspeakers [41, 59]. The considerations regarding reproduction of loudspeakers over headphones are listed below:

1. If the headphones are equalised with respect to frequency response, evaluation of recordings of loudspeakers over headphones is valid for attributes related to timbre, within these specified limitations:

   - Above $\approx 7\,kHz$, the problem of individual differences makes evaluations relying on frequency content in this region unreliable [53], unless individual equalisation is performed.

   - For complex auditory environments, spatial distortion can lead to 'bleeding' of spatial differences onto perceived timbral differences [11], i.e. affect assessor ratings of unrelated perceptual attributes. By spatial distortion is meant changes in perception of spatial attributes, such as reproduction width, depth, localisation, envelopment or similar.

2. For evaluation of certain spatial attributes the use of individual Head-related transfer functions (HRTFs) is required, to obtain sufficient repro-

duction accuracy. With regards to localisation, this is not just a matter of increasing precision, but also of avoiding front-back confusion [58]. Individual HRTFs are needed to correct for both differences between subjects and inter-subject differences, i.e. head- and outer ear asymmetry [57, 58].

3. A final limitation, relates to the physical sensation experienced while listening to stimuli at very high sound pressure levels, especially if the low frequency content is dominating. This is rarely a problem as stimuli is presented at lower SPLs to ensure that assessors do not risk hearing damage from listening sessions. In some cases reproduction of bass content in the presented stimuli may be enhanced by use of a hybrid setup with a set of open headphones and a subwoofer (see e.g. [59]).

Furthermore some concern has been expressed in relation to cross-modal interactions for evaluation of certain stimuli, such as classical concerts. The concern regards the influence on assessors of having tests in small listening booths, compared to the bigger room needed for direct-comparison of loudspeakers, which has a larger visual sense of space. Others have found that it is not a concern for automotive evaluations, where evaluations in-situ in cars, corresponded well to evaluations of recordings conducted in listening booths [11].

7 Conclusions

In this paper a literature study of five aspects of perceptual evaluations were presented with the purpose of maximizing the objectivity of results from perceptual evaluations of loudspeakers.

While the number of sensory descriptors describing perceived dissimilarities between loudspeakers was found in the literature to range from 7 to 55 descriptors, other studies showed that only 2-4 principal dimensions explained a large amount of the variations perceived in loudspeaker evaluations ($87\,\% - 97\,\%$). This is evidence of a low degree of independence among the descriptors, leading to noisy (and inefficient) perceptual evaluations. While the number of sensory descriptors can be decreased if they have a high degree of independence, it is also important to avoid 'bleeding', i.e. all perceptible differences between loudspeaker should be sufficiently covered by the set of sensory descriptors. Finally, it is valuable to consider whether the use of sensory descriptors (words) are sufficient to describe the perceptible differences, or whether e.g. a graphical response format is needed.

Timbre was found in a number of studies looking into perceptual profiling of loudspeakers to be the dominating perceived characteristic between loudspeakers (and headphones) and as a consequence timbral descriptors should

be prioritised in the elicitation process and possibly taken into account in the choice of elicitation methodology. Investigating the models developed by Tan et al. in [28–30] could, to the authors' opinion, be of further value for evaluation of differences of importance in electro-acoustic measurement data on perception.

Differences in perceived loudness of loudspeakers' sound reproduction in listening tests are known to affect sensory ratings significantly and must therefore be minimized. A method by Koehl et al. [40] appeared to be the best loudness equalisation strategy, which introduced individual loudness matching performed by each assessor. This method may potentially be further improved by using a noise signal (white or pink) as loudness matching stimuli (as opposed to the stimuli for the actual listening evaluation), as it was found in [34] that using a white noise stimulus minimized the variability of assessor loudness matchings in a repeated loudness matching evaluation of loudspeakers.

A listening room for perceptual evaluations, should ideally have the reverberation time of typical domestic rooms, for the results to be representative of the average user experience. Furthermore the room response should be flat at mid- and high frequencies to avoid colouration of the acoustical output. Listening rooms complying with the IEC 60268-13 standard was found most suited for evaluation of loudspeakers, while listening rooms complying to the ITU BS.1116-1 standard have too low reverberation times, which may influence the perception of spatial qualities of loudspeakers in stereo or surround setups, and potentially bias the evaluations. With the large current research focus on spatial aspects, this is a needed experimental design variable to consider to maximize measurement accuracy.

Evaluation of loudspeaker recordings presented over headphones was found to be an unbiased alternative to in-situ evaluations; within a set of limitations listed in section 6, setting requirements for equalisation with respect to the nature of evaluation. Most noticeable is the conclusion that individual equalisation is needed for evaluations of any sensory descriptors, affected by frequencies above $\approx 7\,kHz$, i.e. not only evaluation of treble, but also descriptors such timbral balance, clarity, or any other descriptors that may be influenced by the high frequency reproduction. Failing to take this into account could affect accuracy as well as reliability.

Handling these five aspects in perceptual evaluations will improve data objectivity, the prerequisite for making valid scientific conclusions.

Acknowledgement

This work was partly funded by the Danish Agency for Science, Technology and Innovation (Case number: 1355-00061). The authors wish to thank Dan Hoffmeyer, Birgit Rasmussen, Judith Lang, and John LoVerde for helpful

input and data for the Listening room specification section.

References

[1] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, Mass: MIT Press, rev. ed ed., 1997.

[2] F. E. Toole, *Sound reproduction: loudspeakers and rooms*. Amsterdam ; Boston: Elsevier, 2008.

[3] S. Bech and N. Zacharov, *Perceptual audio evaluation: theory, method and application*. Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2006.

[4] S. Bech, "Perception of timbre of reproduced sound in small rooms: Influence of room and loudspeaker position," *J. Audio Eng. Soc*, vol. 42, no. 12, pp. 999–1007, 1994.

[5] S. E. Olive, P. L. Schuck, S. L. Sally, and M. E. Bonneville, "The effects of loudspeaker placement on listener preference ratings," *J. Audio Eng. Soc*, vol. 42, no. 9, pp. 651–669, 1994.

[6] A. Illényi and P. Korpássy, "Correlation between loudness and quality of stereophonic loudspeakers," *Acta Acustica united with Acustica*, vol. 49, pp. 334–345, Dec. 1981.

[7] A. Gabrielsson, "Perceived sound quality of reproductions with different frequency responses and sound levels," *The Journal of the Acoustical Society of America*, vol. 88, no. 3, p. 1359, 1990.

[8] G. Lorho, *Perceived Quality Evaluation, An Application to Sound Reproduction over Headphones*. Ph.d. thesis, Department of Signal Processing and Acoustics, School of Science and Technology, Aalto University, Espoo, Finland, 2010. ISBN 978-952-60-3196-5 (Electronic).

[9] C. Dehlholm, *Descriptive sensory evaluations, Comparison and applicability of novel rapid methodologies*. Ph.d. thesis, University of Copenhagen, Copenhagen, Denmark, 2012. ISBN 978-87-7611-592-0.

[10] A. Gabrielsson, "Perceived sound quality of sound-reproducing systems," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, p. 1019, 1979.

[11] P. Hegarty, S. Choisel, and S. Bech, "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 *Audio Engineering Society Convention 123*, Oct. 2007. Convention Paper 7224.

[12] N. Zacharov, J. Ramsgaard, G. Le Ray, and C. V. Jørgensen, "The multidimensional characterization of active noise cancelation headphone perception," in *Proceedings of the IEEE QoMEX'10 Conference*, (Trondheim, Norway), pp. 130–135, IEEE, June 2010.

[13] S. Olive and T. Welti, "The relationship between perception and measurement of headphone sound quality," in *Audio Engineering Society Convention 133*, Oct. 2012. Convention Paper 8744.

[14] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, p. 968, 2005.

[15] M. Lavandier, S. Meunier, and P. Herzog, "Identification of some perceptual dimensions underlying loudspeaker dissimilarities," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4186–4198, 2008.

[16] S. Choisel and F. Wickelmaier, "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 388–400, 2007.

[17] N. Ford, F. Rumsey, and T. Nind, "Creating a universal graphical assessment language for describing and evaluating spatial attributes of reproduced audio events," in *Audio Engineering Society Convention 115*, Oct. 2003. Convention Paper 5907.

[18] T. Neher, T. Brookes, and F. Rumsey, "A hybrid technique for validating unidimensionality of perceived variation in a spatial auditory stimulus set," *J. Audio Eng. Soc*, vol. 54, no. 4, pp. 259–275, 2006.

[19] H. Stone, J. Sidel, S. Oliver, A. Woolsey, and R. C. Singleton, "Sensory evaluation by quantitative descriptive analysis," *Food Technology*, vol. 28, pp. 24–34, 1974.

[20] J. Pagès, "Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the loire valley," *Food Quality and Preference*, vol. 16, pp. 642–649, Oct. 2005.

[21] B. C. J. Moore, *An introduction to the psychology of hearing*. Leiden: Brill, sixth edition ed., 2013.

[22] R. Bücklein, "The audibility of frequency response irregularities," *J. Audio Eng. Soc*, vol. 29, no. 3, pp. 126–131, 1981.

[23] H. Møller and E. B. Jensen, "On the audibility of non-minimum phase distortion in audio systems," in *Audio Engineering Society Convention 47*, Mar. 1974.

[24] H. Møller, P. Minnaar, S. K. Olesen, F. Christensen, and J. Plogsties, "On the audibility of all-pass phase in electroacoustical transfer functions," *J. Audio Eng. Soc*, vol. 55, no. 3, pp. 113–134, 2007.

[25] S. Choisel and G. Martin, "Audibility of phase response differences in a stereo playback system. part 1: Headphone reproduction," in *Audio Engineering Society Convention 124*, May 2008. Convention Paper 7319.

[26] S. Choisel and G. Martin, "Audibility of phase response differences in a stereo playback system. part 2: Narrow-band stimuli in headphones and loudspeakers," in *Audio Engineering Society Convention 125*, Oct. 2008. Convention Paper 7559.

[27] S. Bech, "Requirements for low-frequency sound reproduction, part i: The audibility of changes in passband amplitude ripple and lower system cutoff frequency and slope," *J. Audio Eng. Soc*, vol. 50, no. 7/8, pp. 564–580, 2002.

[28] C.-T. Tan, B. C. J. Moore, and N. Zacharov, "The effect of nonlinear distortion on the perceived quality of music and speech signals," *J. Audio Eng. Soc*, vol. 51, no. 11, pp. 1012–1031, 2003.

[29] C.-T. Tan, B. C. J. Moore, N. Zacharov, and V.-V. Mattila, "Predicting the perceived quality of nonlinearly distorted music and speech signals," *J. Audio Eng. Soc*, vol. 52, no. 7/8, pp. 699–711, 2004.

[30] B. C. J. Moore, C.-T. Tan, N. Zacharov, and V.-V. Mattila, "Measuring and predicting the perceived quality of music and speech subjected to combined linear and nonlinear distortion," *J. Audio Eng. Soc*, vol. 52, no. 12, pp. 1228–1244, 2004.

[31] ISO, "ISO 28961:2012 acoustics – statistical distribution of hearing thresholds of otologically normal persons in the age range from 18 years to 25 years under free-field listening conditions," Standard ISO 28961:2012, International Organization for Standardization (ISO), 2012.

[32] A. Martini (toim.), "EU work group on genetics of hearing impairment," infoletter 2, European Commission Directorate, Biomedical and Health Research Programme Hereditary Deafness, Epidemiology and Clinical Research (HEAR). EU Work Group 1996, Milano, Italy, Oct. 1996.

[33] W. H. Organization, "Grades of hearing impairment." http://www.who.int/pbd/deafness/hearing_ impairment_grades/en/ [Accessed: July 17th, 2014].

[34] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "The subjective loudness of typical program material," in *Audio Engineering Society Convention 115*, (New York, NY, USA), Audio Engineering Society, Oct. 2003. Convention Paper 5892.

[35] R. M. Aarts, "A comparison of some loudness measures for loudspeaker listening tests," *Journal of the Audio Engineering Society*, vol. 40, pp. 142–146, Mar. 1992.

[36] E. Zwicker and B. Scharf, "A model of loudness summation.," *Psychological Review*, vol. 72, no. 1, pp. 3–26, 1965.

[37] H. Fastl, *Psychoacoustics: facts and models*. No. 22 in Springer series in information sciences, Berlin ; New York: Springer, 3rd. ed ed., 2007.

[38] B. C. J. Moore and B. R. Glasberg, "Modeling binaural loudness," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, p. 1604, 2007.

[39] C. P. Volk and T. H. Pedersen, "System audio - q113. lyttetest med hi-fi-højttalere," Tech. Rep. SenseLab 016/13, DELTA SenseLab, Hørsholm, Denmark, Nov. 2013. Source in Danish.

[40] V. Koehl and M. Paquier, "Influence of level setting on loudspeaker preference ratings," in *Audio Engineering Society Convention 126*, (Munich, Germany), May 2009. Convention Paper 7782.

[41] V. Koehl, M. Paquier, and S. Delikaris-Manias, "Comparison of subjective assessments obtained from listening tests through headphones and loudspeaker setups," in *Audio Engineering Society Convention 131*, Oct. 2011. Convention Paper 8560.

[42] W. Geddes, T. Somerville, C. Gilford, and A. Newman, "The influence of listening conditions on the quality of reproduced speech," Research paper Report No. B.060 (Serial No. 1954/27), BBC Research & Development, Salford, UK, June 1954.

[43] ITU-R, "Recommendation BS 1116-1, methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.," Recommendation ITU-R BS 1116-1, International Telecommunication Union Radiocommunication Assembly (ITU-R), United States, 1997.

[44] IEC, "Sound system equipment. 13. listening tests on loudspeakers," Recommendation IEC 60268-13, International Electrotechnical Commission (IEC), Geneva, 1998. 2nd edition.

[45] AES, "AES recommended practice for professional audio - subjective evaluation of loudspeakers (reaffirmed 2007)," Standard AES20-1996 (r2007), Audio Engineering Society, New York, NY, USA, 2007.

[46] EBU, "Listening conditions for the assessment of sound programme material: monophonic and two-channel stereophonic," Technical document EBU Tech. 3276 - 2nd edition, European Broadcasting Union, Geneva, Switzerland, May 1998.

[47] EBU, "Listening conditions for the assessment of sound programme material: Multichannel sound (supplement 1)," Technical document EBU Tech. 3276-E, European Broadcasting Union, Geneva, Switzerland, May 2004.

[48] G. Jackson and H. Leventhall, "The acoustics of domestic rooms," *Applied Acoustics*, vol. 5, pp. 265–277, 1972.

[49] M. Burgess and W. Utley, "Reverberation times in british living rooms," *Applied Acoustics*, vol. 18, pp. 369–380, 1985.

[50] C. Díaz and A. Pedrero, "The reverberation time of furnished rooms in dwellings," *Applied Acoustics*, vol. 66, pp. 945–956, Aug. 2005.

[51] J. Lang, "Data: Reverberation times in domestic furnished rooms in austria," 2012. Judith Lang data from colleagues in Salzburg and Vienna, 2011, established for discussion in COST TU0901.

[52] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention—focusing the searchlight on sound," *Current Opinion in Neurobiology*, vol. 17, pp. 437–455, Aug. 2007.

[53] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc*, vol. 43, no. 4, pp. 203–217, 1995.

[54] J. Merimaa, "Modification of HRTF filters to reduce timbral effects in binaural synthesis," in *Audio Engineering Society Convention 127*, Oct. 2009. Convention paper 7912.

[55] J. Merimaa, "Modification of HRTF filters to reduce timbral effects in binaural synthesis, part 2: Individual HRTFs," in *Audio Engineering Society Convention 129*, Nov. 2010. Convention paper 8265.

[56] L. A. Gedemer and T. Welti, "Validation of the binaural room scanning method for cinema audio research," in *Audio Engineering Society 135th Convension*, (New York, NY, USA), Audio Engineering Society, Oct. 2013. Convention Paper 8974.

[57] J. Merimaa, V. R. Algazi, and R. O. Duda, "Individual perception of headphone reproduction asymmetry," in *Audio Engineering Society Convention 131*, Oct. 2011. Convention Paper 8540.

[58] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc*, vol. 44, no. 6, pp. 451–469, 1996.

[59] A. Lindau and F. Brinkmann, "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings," *J. Audio Eng. Soc*, vol. 60, no. 1/2, pp. 54–62, 2012.

Paper B

Identifying the dominating perceptual differences in headphone reproduction

Christer P. Volk, Mathieu Lavandier,
Søren Bech, and Flemming Christensen

The article has been submitted and is currently in the second review process for publication in the
Journal of the Acoustical Society of America, pp. 1–12, 2016.

To accommodate the smaller format of the printed thesis publication, the layout has been revised. The text is identical to the submitted/published article.

Abstract

The perceptual differences between the sound reproductions of headphones were investigated in a pair-wise comparison study. Two musical excerpts were reproduced over 21 headphones positioned on a mannequin and recorded. The recordings were then processed and reproduced over one set of headphones to listeners, who were asked to evaluate their perceived degree of dissimilarity. The two musical excerpts were used in separate experiments. The processing of the recordings consisted of compensating for the influences of the playback headphones worn by the listeners as well as for the mannequin's ear canals. A multidimensional scaling analysis revealed two dominating perceptual dimensions used by the listeners to differentiate the reproductions of the headphones. These dimensions were similar for the two musical excerpts. Objective metrics are proposed to describe them, leading to correlations ranging from 0.89 to 0.97 between the dimensions and metrics. The first perceptual dimension was associated with the relative strength of bass, while the second dimension was related to the relative strength of the lower midrange.

1 Introduction

Listening over headphones has gained increasing popularity during the last decade or more and with it the interest in making better headphones. The headphones research effort however extends much further back in time and has had several focus areas, such as the reduction of distortion [1], the choice of the design target for the frequency response (see e.g. [2–4]), as well as the perceptual investigations of general sound quality [5], subjective preference [6], spatial characteristics [7, 8], and the preservation of cues important for externalisation [9]. Furthermore, investigations into the characteristics of headphones have been conducted in an effort to differentiate between the numerous available headphone models [8, 10], as understanding of the perceptually dominating characteristics may focus the research effort towards the limiting factors of audio reproduction. The aim of the present study was to highlight the dominating auditory perceptual dimensions differentiating headphones and to relate these dimensions to objective metrics.

In the recent years, a number of studies have focused on quantifying the optimum headphone reproduction leading to the highest preference among listeners [5, 6, 11]. These studies have relied on the assumption that a global preference exists, a "one size fits all". It has however been established that clusters exist with regards to listener preference in a wide range of applications and areas, e.g. binaural mixing algorithms [12], multichannel audio quality [13] or other sensory domains such as tasting [14]. Since audio reproduction free from influences from the reproduction equipment is not possible, compromises must be made in the electro-acoustic design of headphones.

Manufacturers must prioritize between these compromises in the effort to minimize the difference between the design target and the obtained audio reproduction. This has the consequence that, even if all listeners preferred the same ideal (the manufacturers design target), they might not prefer the same compromise. This issue can be investigated by performing a perceptual characterisation of headphones and study the relationship between the obtained characteristics and the observed clustering in listener preference among consumers. The efforts of perceptual characterisation of headphones [6–8, 11] have led to a large number of sensory descriptors[1] describing the differences between headphones. It is however unclear how important the individual sensory descriptors are for the overall sound perception of headphones.

[7] tested eight headphones. Twenty listeners evaluated the similarity of the headphone reproduction of five musical excerpts on 30 sensory descriptors (adjectives). A factor analysis of the data led to five main factors needed for differentiating the headphones: sharpness/hardness, clearness/distinctness, disturbing sounds, brightness/darkness, and feeling of space. The study gave insight into characteristics differentiating headphones for the included set of headphones, but had a number of limitations. The 30 sensory descriptors were imposed on the listeners, as the original sensory descriptor elicitation process was conducted without the listeners participating in the study. Consequently all perceivable differences might not have been described and the listeners, naïve to sound reproduction evaluation, might not have been able to fully understand the proposed descriptors. Furthermore, the reproduction technologies of the included headphones are no longer representative of the products available today, which are almost exclusively electro-dynamic transducers. Additionally, the listeners were inexperienced with headphone listening, which may also have affected their expectations. Altogether the conclusions of the study may not have been representative of differences between headphones in general at the time, due to the limited number of products, and may not be representative of perceived headphone differences today, due to the shift in technology and the increased usage of headphones for listening to music.

In [6], listener preference among six headphones was evaluated. Ten listeners were asked to describe the characteristics of each set of headphones, leading to 19 sensory descriptors. Finally, the frequency of occurrence of each of these 19 terms was analysed with regards to correlation with the preference scores, thereby describing the main characteristics important for preference ratings (both negative or positive effects). The three most positively correlated descriptors were: good spectral balance, wide sound stage and neutral/low coloration, while the three descriptors leading to the low-

[1] A sensory descriptor is defined here as a word or phrase that describes, identifies, or labels a perceptual characteristic of a system, e.g. a headphone reproduction. This definition is adapted from [15].

est negative correlations were: distorted, dull and coloured. While providing valuable insights regarding the factors influencing preference ratings of the six tested headphones, the listeners were specifically trained to evaluate linear and non-linear distortion, which may have affected their focus with regards to preference ratings of headphones.

In [6, 7, 11], the headphone evaluations were performed as a real-device test, i.e. with the listeners wearing the headphones during the evaluations. This has the advantage of getting an accurate representation of the headphones' sound reproduction, but can also lead to biases related to other factors than the reproduction, such as the sense of comfort [6]. Furthermore, direct comparison of headphones is affected by the time needed to switch between headphones, increasing the complexity of the listeners' evaluation task and making the data quality very dependent on their ability to memorise the characteristics of one headphone while having another put on. In these three studies, the headphones were put on by a test instructor ("semi blind" paradigm) to limit biases related to listeners' interaction with the headphones. This may however also influence the evaluations, e.g. by leading to a non-optimum fit on the listeners ears or by influencing the listeners by being present during the test.

In [8], six headphones (five with active noise cancellation (ANC) technology) were evaluated on ten descriptors. The reproductions of the headphones were recorded before being compared. The recordings were made with the headphones placed on a mannequin with background noise playing from external loudspeakers and either with or without a musical excerpt being reproduced by the headphones themselves. The stimuli, representing the six headphones, were then reproduced over a pair of reference headphones in the listening test, thereby allowing instantaneous comparison in a double-blind paradigm. The test with the musical excerpt led to four main dimensions identified: timbre/attenuation, dynamic/spatial, precision/stereo space, as well as treble range. The paradigm led to an efficient evaluation of ANC headphones, with the compromise of using recordings to represent the headphones.

Lavandier and collaborators developed a protocol to investigate perceptual dimensions used to differentiate the sound reproduction of loudspeakers in a monophonic setup [16–18]. The reproduction of musical excerpts reproduced over a large number of loudspeakers was recorded and later presented over headphones to listeners in pair-wise comparisons. This protocol led to stable results for a large collection of loudspeakers [18] and for different musical excerpts [16, 17]. The result of a multidimensional scaling (MDS) analysis of the data collected by [18] revealed three main perceptual dimensions: bass/treble balance, relative strength of midrange, and feeling of space. It would be of interest to investigate the influence of differences in audio reproduction separately from room influences, as is possible in evaluation of

headphone reproduction.

Lavandier and collaborators showed that many reproduction systems are needed to uncover all perceptual dimensions: three dimensions were uncovered in the study with 37 loudspeakers [18], while only the first two were found in the previous studies with 12 loudspeakers [16, 17]. The previously mentioned headphone characterisation studies only included between 4 and 8 headphone models [6–8, 11]. Additionally, when one wants to describe the perceptual dimensions with objective metrics, involving more stimuli increases the chances of having the stimuli more homogeneously spread along the different perceptual dimensions, allowing for a more reliable objective description of these dimensions. Finally, comparing more loudspeakers or headphones should lead to perceptual dimensions, which are more representative of loudspeakers/headphones in general rather than of the particular set of systems under test.

The main purpose of this study was to investigate the extent to which the results with loudspeakers were valid for headphones, i.e. for audio reproduction with neither room influences nor artefacts from systems with multiple drivers. By including a much larger set of headphones than presented in previous studies, the aim was to establish a more general understanding of the main characteristics of headphones audio reproduction and the perceptual significance of measurable differences. Additionally, it was investigated whether an optimization routine, would lead to metrics describing the perceptual differences with higher correlations than previously found in the literature.

The present study consisted of two listening tests, involving different musical excerpts, but the same experimental protocol and the same 21 headphones. As in the protocols proposed by [16, 17] and [8], recordings of the headphones were used as an alternative to real-device evaluation to avoid problems related to differences in interaction between individuals and headphones [19] as well as removing biases related to other modalities, such as visual or haptic impressions. Naïve listeners were instructed to evaluate the degree of dissimilarity between the headphones in a pair-wise comparison paradigm similar to that of [17]. The pair-wise comparison data comprises a direct measure of dissimilarity, which provides insight into the latent structure of the listeners internal decision criteria. An MDS analysis was used to describe this structure and reveal its dimensionality. Metrics are proposed to describe the dominating perceptual dimensions. These metrics are based on computations of the spectral content of the test stimuli, but the estimated loudness spectrum was used rather than the frequency spectrum, to take the properties of the human auditory system into account (e.g. middle ear influence, frequency-dependent sound level perception, and frequency masking).

Table B.1: Distribution of the type of headphones included in this study.

	Open-back or Semi-open-back	Closed-back
Circumaural	4	9
Supraaural	1	7

2 Methods

2.1 Headphones

A set of 21 electrodynamic headphones was included in each experiment. It consisted of six prototypes from one manufacturer, nine commercially available models from another manufacturer, and additional models from six other manufacturers. The models were a mix of open- and closed-back headphones, circumaural and supraaural models as shown in Table B.1, and spanning more than a factor 10 in price (from $\approx$ 30 USD to $>$ 350 USD).

2.2 Recording and processing

The stimuli for each experiment consisted of the recordings of the sound reproductions from the 21 headphones. These recordings were reproduced over a pair of Sennheiser HD 650 headphones (referred to as the playback headphones in the following). The processing steps utilised to record and prepare the stimuli in the experiments were the following. Musical excerpts were played back over the headphones positioned on a B&K HATS 4128C mannequin and the binaural output was recorded in 32 bit/48 kHz WAV format. The headphone positioning was checked by recording pink noise and comparing the right-left input level balance prior to recording of the musical excerpts. In cases of leakage, the level was clearly lower in one channel. In cases with imbalanced driver sensitivity or poor mannequin ear fit, perceivable left-right imbalance could not be completely avoided. The recording gain was adjusted for each set of headphones to obtain similar output levels independent of their sensitivity. A control listening concluded each gain adjustment to avoid a playback level with increased non-linear distortion.

The influence of the ear canals of the mannequin was removed with an 128th order minimum-phase inverse finite impulse response filter based on measurements by B&K of the ear canal influence from the specific mannequin. The influence of the playback headphones were compensated by an inverse filter designed using a MATLAB toolbox in development [20]. The toolbox facilitates the design of minimum-phase, linear-phase and zero phase-filters. A minimum-phase filter was chosen to compensate for the am-

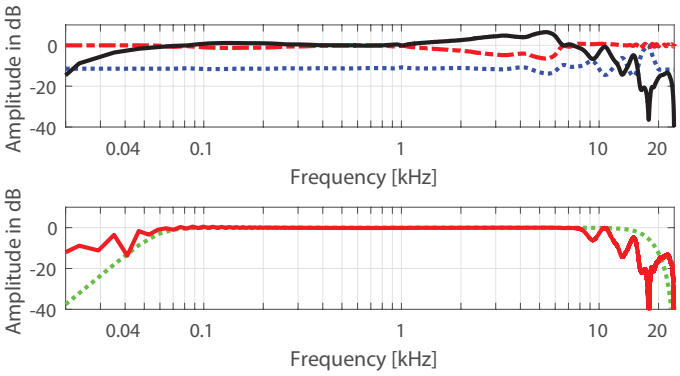


Fig. B.1: (Color online) Inverse filter design. The upper plot shows the measured mean transfer function (line), the regularisation curve (dotted) and the regularised inversion target (dash-dotted). The lower plot shows the target inverse filter (dotted) and the compensated transfer function (line).

plitude response of the playback headphones while avoiding reported potential pre-ringing artefacts of linear-phase filters [21]. The filter design included averaging over measurements, regularisation, octave smoothing of the measured transfer function, and 1/9-octave smoothing of the regularisation. These parameter settings in the toolbox were based on informal experimentation and perceptual evaluations by two of the authors. The inverse filter was based on measurements of the frequency response of nine Sennheiser HD 650 headphones positioned on the mannequin and averaged across channels. Plots of the playback headphones frequency response and compensation filter are shown in Figure B.1. The resulting filter was irregular in the lowest and highest frequencies, but had an extended bass response, which improved the perceived transparency for the musical excerpts compared to a smooth and regular filter with less bass extension.

Informal perceptual loudness equalisation by two normal-hearing listeners was conducted. Using the playback headphones, the perceived loudness of each processed recording were evaluated relative to the unprocessed original musical track reproduced at the chosen playback level of the experiments. In the final processing step, the excerpts were converted to 16 bit WAV files to obtain compatibility with the test software.

2.3 Stimuli

The musical excerpts used for the headphones recordings originated from an electronic music track by Todd Terje (*Delorean Dynamic (Disco Mix)*. Album: It's Album Time. 2014. Olsen Records, Norway) and a soft pop track by Tina Dickow (*Room with a view*. Album: In the Red. 2006. Finest Gramo-

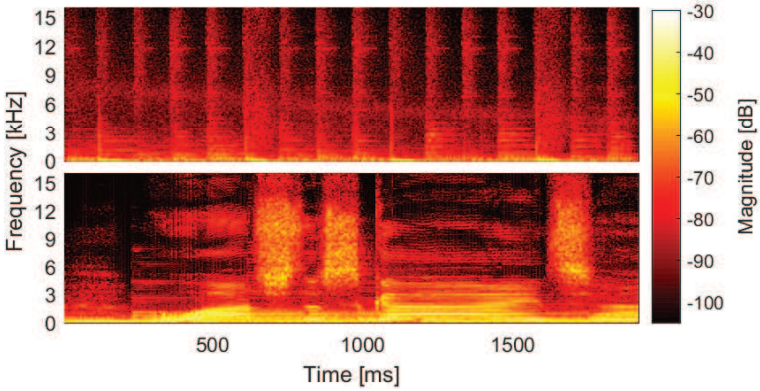


Fig. B.2: (Color online) Spectrogram of the two musical excerpts: Todd Terje (top) and Tina Dickow (bottom). Averaged over the two channels. Note: Only the first 1.9 s of the Tina Dickow excerpt is shown.

phone/A:larm/Universal Music, London UK). These two excerpts were chosen among nine excerpts covering a wide range of genres. By informal listening, excerpts were selected which provided the highest discrimination between headphones. Short samples of 1.9 and 4.5 seconds, respectively, were selected from the full tracks, the samples being representative of the tracks and cropped to maintain the rhythm during their looping allowing listeners sufficient opportunity to get a good impression of each stimulus. The limited auditory memory of humans [22] makes it attractive to use stimuli of short durations in auditory comparison tasks. For listening tests with trained listeners and small differences between stimuli, longer stimuli durations can be of value because it allows the experienced listeners to make their own choice of the best part of the excerpts for discrimination. In this study, involving naïve listeners, it was considered better to pre-select a discriminating part of the excerpts to maximize discrimination and stability in the evaluations. Samples of such short durations were also used in [17] and led to meaningful MDS spaces, which were stable across samples. Short samples ensured that all listeners base their judgement on the same part of the original excerpt.

Spectrograms of the two excerpts are plotted in Figure B.2. The Todd Terje excerpt consists of a fast electronic beat with a deep bass and a wide band spectrum and the Tina Dickow excerpt consists of a short sentence (*"Watch the sky turn from hazy grey to black"*) accompanied by a prominent acoustic guitar. It has an emphasized midrange and some full band 's' sounds. Both excerpts make little use of stereo effects, i.e. having a largely centred stereo image.

As described in Section II.B the recordings were post-processed in an effort to compensate for the effects of the playback headphones and the man-

nequin. In order to investigate the precision and influence of this auralization process, an additional stimulus was included in each experiment: the original unprocessed musical excerpt. When reproduced over the playback headphones this stimulus could be compared with an auralized version of the same playback headphone model also reproduced over the playback headphones (as the Sennheiser HD 650 were one of the 21 headphones under test). The mean perceived difference between these two stimuli was used as an indicator of how suited the recording and processing steps were for the headphones evaluation (see the 4).

The playback level was adjusted to a listening level considered comfortable for a 90-minute listening test, selected by informal listening by two listeners. This level corresponded to $L_{eq} = 72 \pm 2\,dB\,SPL$ for the two unprocessed musical excerpts adjusted to the same loudness as the remaining stimuli.

2.4 Listening test procedure

The experiments consisted of a pair-wise comparison task with French speaking listeners. Listeners were instructed to evaluate the dissimilarity between pairs of recordings involving two different headphones (but the same musical excerpt). The comparison scheme was a half-matrix combination of the stimuli, e.g. comparison of headphone A vs. B but not B vs. A nor A vs. A. This led to 231 comparisons for 22 stimuli. The user interface had a $16\,cm$ continuous and unipolar response scale with the verbal anchors "Pas du tout différent" (not different at all) and "Extrêmement différent" (extremely different) positioned at the two extremes of the scale, as well as a tick at the center of the scale. Listeners rated the perceived difference between stimulus A and B by moving a marker on this horizontal rating scale.

The listening tests were conducted in a double-walled soundproof booth. A computer screen, visible to the listeners, was placed outside the booth, while the keyboard and mouse were placed inside the booth. Digital/analog conversion and amplification of the stimuli were accomplished using a Lynx TWO sound card. Stimuli were presented in randomized order via the playback headphones.

Listeners completed two short familiarisation tasks prior to the dissimilarity evaluation. First, a task with informal listening to 12 stimuli from the test, selected to be representative of the range of differences in the stimuli. Listeners were required to listen at least once to each stimulus. Secondly, a task with a minimum of 10 pair-wise evaluations (using randomly-chosen pairs from the dissimilarity evaluation) to familiarise the listeners with the pair-wise comparison methodology and the user interface. Their use of the software was monitored from outside the listening booth to ensure correct use and understanding of the task. Before starting the dissimilarity evaluation listeners were encouraged to take at least one break during the experiment.

Each listening test lasted about 1-1.5 hours.

After each experiment listeners were asked if they could put words on the type of differences they heard. Their responses were useful for the later process of establishing hypotheses regarding the nature of the dominating perceptual differences and for the evaluation of potential biases in the test (see the 4).

2.5 Listeners

Fifteen naïve listeners participated in each experiment. Eight listeners from the first experiment participated in the second experiment. The Todd Terje experiment was conducted prior to the Tina Dickow experiment with approximately 3 weeks between sessions for listeners participating in both experiments. In both experiments, 7 men and 8 women participated. The age of the listeners was not normal-distributed, but ranged from 19 to 31 with a median of 22 in the Todd Terje experiment. The age distribution was similar in the Tina Dickow experiment, but with a minimum age of 20. The participants were paid for their participation. All participants reported having normal-hearing, but were not screened for hearing loss. After the listening tests, hearing loss was discovered in one listener, but the listener's ratings were kept in the dataset, as they were not found to be outliers.

The number of listeners included in the two tests was based on monitoring of the change in the average difference in the dissimilarity half-matrix with the addition of ratings from each new listener. As a result, the number of listeners needed to get a stable average could be monitored and justified prior to analysis of the results. The average difference between dissimilarity matrices with N subjects vs. N-1 subjects is depicted as a function of N in Figure B.3. Average difference is defined as the summed differences of ratings divided by the number of ratings in the half-matrix. This average difference is scaled to the response scale in the experiments, i.e. $0 - 16\,cm$. Figure B.3 shows that a stable level was reached after the 12th listener contribution for both musical excerpts. It is seen that the average change approaches an asymptote for 15 listeners, where ratings of dissimilarity between headphones are stable within approximately $\pm 0.4\,cm$ on average.

2.6 Analysis method

The dissimilarity data were analysed using a metric MDS analysis [23]. The method is based on the Classical MDS [24], but uses Euclidean distance calculations. The MDS algorithm calculates the N-dimensional space in which the distance between all pairs are as close as possible to the dissimilarities contained in the perceptual dissimilarity matrix. Following the hypothesis underlying the MDS analysis, the MDS dimensions represent the criteria used

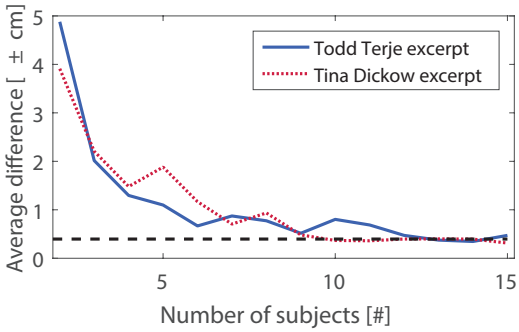


Fig. B.3: (Color online) Monitoring of the average difference in dissimilarity matrices with addition of ratings from each new listener. Average difference is defined as the summed differences of ratings between the matrices with N vs. N-1 subjects divided by the number of ratings in the half-matrix. The dotted line marks the asymptote at $\pm 0.4\,cm$, which is the average difference for 12 to 15 listeners.

by the listeners to evaluate dissimilarity — although the meaning of the dimensions is a matter of interpretation. The algorithm can be constrained to find the optimum solution for any choice of N. While the relative distance between all stimuli in the MDS space are fixed, the scaling and rotation of the whole MDS space are unconstrained. Typically, the space is rotated, such that the axes present the MDS dimensions in monotonically decreasing order of variance, based on the eigenvalues of the MDS space. Other rotations may however be of value for the interpretation of the space.

The metric MDS method is based on the premise that all listeners use identical internal rating criteria in the pairwise comparison evaluations. This includes having the same perception of the stimuli, basing the decision on the same perceived characteristics and giving the same weights to each of these perceived characteristics. This is a simplification of real life conditions, but has the strength that no structure is imposed on the data, and only fundamental dimensions, shared by the majority of listeners, will be prominent in the resulting multidimensional space. The number of important dimensions in an MDS analysis, i.e. dimensions related to perception and not to random noise in the evaluations, is often evaluated using the STRESS metric. It is based on a cost function, which measures the similarity between the MDS configuration, e.g. a 2-dimensional MDS, and the dissimilarity matrix (raw data). When plotting the STRESS calculated for each number of dimensions a knee-point is often clear, which is used as a dimension-selection criteria. The implication of the knee-point is that adding dimensions beyond this point will increase the similarity with the dissimilarity matrix only a little, but it is likely to additionally add noise and complicate the interpretation of the MDS space.

In pairwise comparison experiments the number of independent (orthogonal) dimensions possible to uncover in the analysis depends on the number of stimuli included in the test. A literature study [25] showed that up to 4 or 5 independent dimensions are likely to exist concerning the perceived characteristics of headphones. A rule-of-thumb [26] states that the number of stimuli N needed to statistically uncover N_{dim} dimensions can be estimated by this equation: $N_{dim} = \frac{(N-1)}{4}$. Consequently, an estimated minimum of 17 stimuli is needed to uncover 4 dimensions and 21 stimuli for 5 dimensions. With 21 headphones (+ the original sample) included in the present experiments, approximately 5 dimensions could potentially be uncovered in the MDS analysis.

3 Results

3.1 MDS analysis

For both experiments, plotting the STRESS metric as a function of the number of MDS dimensions displayed a knee-point at two dimensions. Furthermore, a bootstrapping procedure with 500 iterations was used to estimate the 95 % confidence intervals between headphones in the third dimension of a 3-dimensional MDS space. This procedure revealed no significant differences between headphones in the third dimension. As a result the 2-dimensional MDS space was selected for further analysis. In Figure B.4 and Figure B.5 the resulting 2-dimensional MDS maps are depicted. Each point refers to a stimulus/headphone. The MDS map for the Tina Dickow experiment has been rotated to best match the dimensions for the Todd Terje experiment (arbitrarily chosen as a reference to compare the two spaces) using a generalized procrustes analysis procedure. Confidence intervals are shown for the unprocessed musical excerpt (Original) and the auralized Playback headphones for validation of the processing steps (see the 4). Note that most headphones are positioned similarly within the space regardless of the experiment's musical excerpt.

Using the MDS maps, structured listening to the stimuli of each experiment was conducted by two of the authors in an effort to establish hypotheses regarding the underlying nature of the two dimensions. Dimension 1 was perceived as clearly related to bass, while dimension 2 was perceived as related to treble or midrange/treble ratio. Both dimensions were thus hypothesized to be spectral in nature. As spectral dimensions were also found as two of the dominating dimensions in the previous work by Lavandier and collaborators [16–18], the modelling effort presented in the next section investigated only this aspect of audio reproduction.

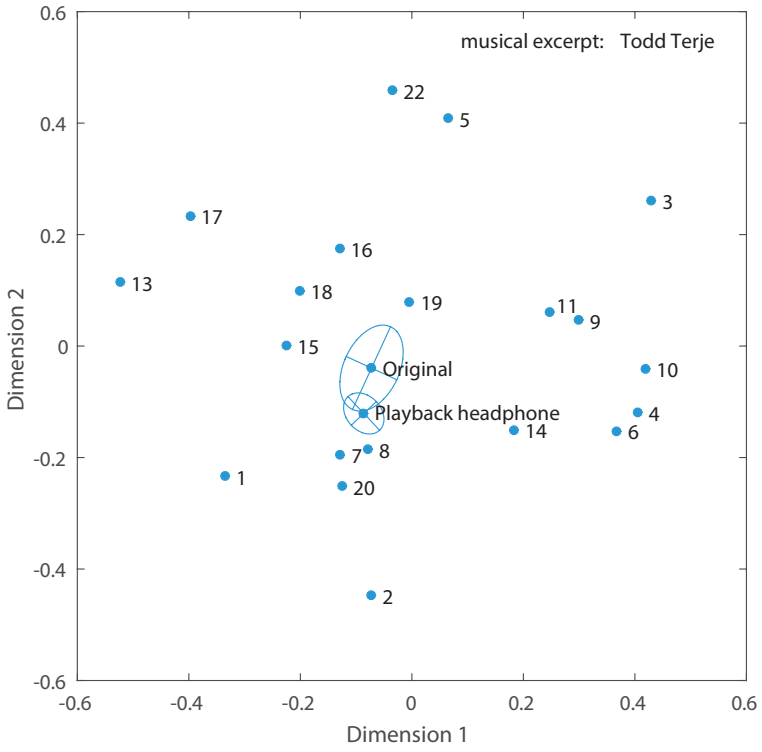


Fig. B.4: (Color online) Two-dimensional MDS map resulting from the dissimilarities evaluated in the experiment with the Todd Terje excerpt. Each point refers to an auralized headphone. 'Original' refers to the original unprocessed excerpt and 'Playback headphone' refers to the recorded and processed pair of Sennheiser HD 650 played back over the HD 650 used in the experiment. The ellipses are the 95% confidence ellipses of the stimuli used to discuss the validation of the stimuli processing (see the 4).

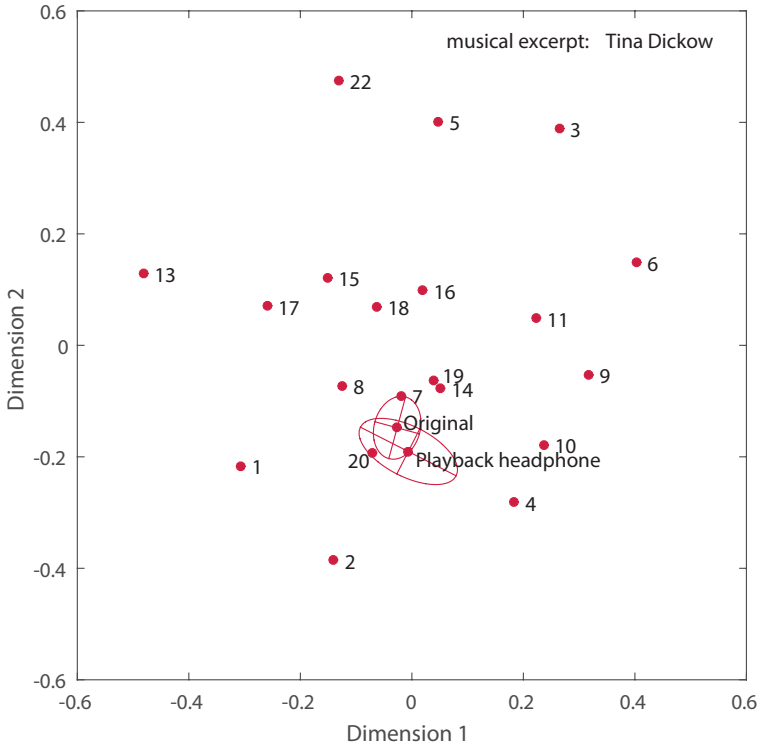


Fig. B.5: (Color online) Same as Fig. B.4 but for the experiment with the Tina Dickow excerpt.

3.2 Link between MDS dimensions and stimuli

Metrics were developed to investigate the link between MDS dimensions and stimuli. These metrics were all based on a time-varying loudness model [27], where the time-averaged specific instantaneous loudness was calculated for each stimulus in the experiments in steps of 0.25 Equivalent Rectangular Band (ERB) from 0.8 ERB (20 Hz) to 38.8 ERBs (14.7 kHz). The loudness metrics were all constructed as the summed loudness in a frequency range AB divided by either the full loudness spectrum (FS) or the loudness in another non-overlapping frequency range CD.

An optimisation routine was utilised to find optimum areas AB and CD at which the metrics have the maximum correlation with the coordinates in MDS dimension 1 or 2. As it was not of interest to get metrics related specifically to each of the two musical excerpts, but rather to get metrics of a more general nature, the metrics utilizes the frequency ranges (AB and CD) with the maximum average correlation across the two musical excerpts. An outline of the optimisation routine is shown in Figure B.6. All combinations of search ranges and search positions were tested with two constraints: 1) the search ranges were restricted to a minimum width of two ERBs, to avoid getting maximum correlations in too narrow frequency ranges unlikely to be the cause of the perceptual evaluation and 2) the two frequency ranges AB and CD were not allowed to overlap. The optimisation approach was selected to test the hypotheses established by informal listening, i.e. that the two perceptual dimensions were related to one frequency range relative to another range (including full-range). To allow an outcome leading to a rejection of the hypotheses the search ranges AB and CD included the full loudness spectrum.

The metrics were calculated for both the left and right channel loudnesses as well as for their average in each 0.25 ERB (Ch. average) and the largest of the two loudnesses in each 0.25 ERB (Ch. max). The two latter cases combining the information from both channels were included to reduce the influence of any specific characteristic in the stereo-mix of the chosen musical excerpts. 'Ch. average' was included to represent a binaural summation strategy, while 'Ch. max' may represent a 'better ear' strategy. Note that the time-varying loudness model simulates the spreading of excitation between successive time frames and consequently better estimates the time-averaged loudness. The equations for the suggested metrics are presented in Eq. B.1 to B.4. Since the "Ch. average" approach led to the best results in terms of correlation with the perceptual dimensions, the frequency limits in the equations below are reported for this data basis only (they were similar for the left, right and "Ch. max" loudnesses).

Two metrics gave almost identical maximum correlations with the perceptual dimension 1. The equations for these two metrics *DeepBass* and

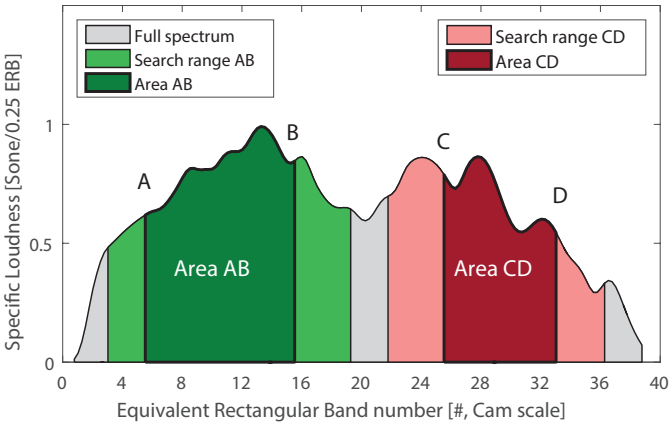


Fig. B.6: (Color online) Sketch of the optimisation routine used to define the metrics on the specific loudness of the stimuli. The lightest area represents the full frequency range. The two dark areas represent examples of frequency ranges investigated in one iteration of the routine within the two respective search areas. The letters A-D denote the start- and end-points of the dark areas of this iteration.

Bass / Mid are given in Eq. B.1 and Eq. B.2 respectively, with f being the frequency and $Dens_m$ being the temporal mean of the time-varying specific loudness. Note that an empty summation sign denotes a summation over the full loudness spectrum.

$$DeepBass = \frac{\sum_{f=20\,Hz}^{186\,Hz} Dens_m(f)}{\sum Dens_m} \tag{B.1}$$

DeepBass is the ratio between the specific loudness in the frequency range $20 - 186\,Hz$ ($ERB_N = 0.8 - 5.8$) and the loudness over the full spectrum ($ERB_N = 0.8 - 38.8$).

$$Bass/Mid = \frac{\sum_{f=20\,Hz}^{222\,Hz} Dens_m(c)}{\sum_{f=246\,Hz}^{3262\,Hz} Dens_m(c)} \tag{B.2}$$

Bass / Mid is the ratio between the specific loudness in the deep and low bass $20 - 222\,Hz$ ($ERB_N = 0.8 - 6.3$) and the specific loudness in the midrange $246\,Hz - 3262\,Hz$ ($ERB_N = 6.8 - 25.3$).

Two metrics, well-correlated with the perceptual dimension 2, are de-

scribed by Eq. B.3 and Eq. B.4.

$$Bass{+}Mid = \frac{\sum_{f=144\,Hz}^{2039\,Hz} Dens_m(f)}{\sum Dens_m} \tag{B.3}$$

Bass+Mid is the ratio between the specific loudness in the bass and lower midrange $144 - 2039\,Hz$ ($ERB_N = 4.5 - 21.3$) and the loudness in the full spectrum.

$$Mid/Treble = \frac{\sum_{f=330\,Hz}^{1369\,Hz} Dens_m(f)}{\sum_{f=3659\,Hz}^{13940\,Hz} Dens_m(f)} \tag{B.4}$$

Mid/Treble is the ratio between the specific loudness in the high bass and low midrange $330 - 1369\,Hz$ ($ERB_N = 11.3 - 16.0$) and the specific loudness in the treble $3659 - 13940\,Hz$ ($ERB_N = 30.3 - 38.0$).

The maximum correlations resulting from the optimization routine are shown in Table B.2. For the metric *DeepBass* the sign of the correlation is different for the left and right channels. This is caused by the optimization routine leading to two peaks with almost identical maxima. One peak related to bass vs. full-range and the other peak related to everything-but-bass vs. full-range (the reciprocal). The two peaks thus describe the same: the relative level of the bass. These two peaks are found for *DeepBass* regardless of the data basis (Left, Right, Ch. average or Ch. max). An example of the two peaks are shown in Figure B.7. From Table B.2 it is also seen that, on average across musical excerpts, the highest correlation is always found for the column 'Ch. average'. Consequently only the 'Ch. average' correlations are discussed hereafter. Note also that the difference in correlation values between the two excerpts for Bass+Mid and Mid/Treble, in combination with 'Right ch.', are caused by one set of headphones being an outlier. Without this outlier, correlations are $r = 0.89$ and $r = 0.87$ for *Bass/Mid* and *Mid/Treble* respectively."

The frequency ranges used in the equations of the four metrics are based on the maximum correlations found for the current dataset, but for prediction of other headphones/musical excerpts the relevant frequency ranges could shift, expand or contract. In Figure B.7 the correlation between *DeepBass* and perceptual dimension 1 is plotted as a function of the start frequency A and end frequency B of the AB range used to compute the metric (i.e. the output of the optimization routine). As mentioned previously two peaks were found for *DeepBass*, with one being the reciprocal of the other. In this case it was hypothesized that *DeepBass* would be more stable than its high frequency counterpart, as it has a lower dependence on the uncertainties at

Table B.2: Pearson correlation coefficients between loudness metrics and MDS dimensions. The loudness metrics are specified in the rows and the data basis (calculated loudness for one or two channels) in the columns. The two numbers in each cell represents the correlation for the Todd Terje and Tina Dickow excerpts respectively. All correlations are significant on a $\alpha = 0.001$ level.

Metric	Dim.	Left ch.	Right ch.	Ch. average	Ch. max
DeepBass	1	0.90/0.86	-0.95/-0.94	0.97/0.94	0.92/0.89
Bass/Mid	1	0.89/0.88	0.95/0.93	0.97/0.94	0.92/0.88
Bass+Mid	2	0.92/0.95	0.76/0.94	0.88/0.97	0.85/0.95
Mid/Treble	2	0.91/0.91	0.75/0.94	0.89/0.95	0.85/0.94

high frequencies, where physical differences between listeners' ears causes the signal reaching the ear drum to vary [28]. With regards to the uncertainty of the limits of the frequency ranges of the *DeepBass* metric, it was found that the correlations were within 0.04 of the maximum correlation within a shift of A or B by plus or minus 2-3 ERBs. For larger changes in the frequency limits, correlation between *DeepBass* and dimension 1 drops more rapidly. A similar sensitivity was observed in the definition of the frequency limits used for the other metrics of this study.

4 Discussion

Judging from the similarity of the MDS maps for the two musical excerpts (Fig. B.4 and B.5), the same internal criteria were used for evaluation of the differences between headphones in the two experiments, despite the difference in musical genre (electronic beat vs. soft pop) and the corresponding differences in temporal and spectral characteristics of the excerpts being reproduced. While the same dimensions were found for both excerpts, the ranking based on the percentage of variance explained by each dimension interchanged from one excerpt to the other (D1: 32 % and D2: 16 %; D1: 21 % and D2: 24 % for the Todd Terje and Tina Dickow experiments respectively), suggesting that the perceptual weight of each dimension depended on the stimuli. Note that some debate exists regarding the method of calculating explained variance for MDS dimensions, so the weighting of the separate dimensions should be considered here with caution. In [16], an experiment with pairwise comparison of loudspeakers included three musical excerpts of different genre. An MDS analysis of the dissimilarity data led to very similar MDS 2-dimensional spaces across excerpts, supporting the findings of the current study that perceptual dimensions can be stable across musical excerpts.

With regards to the metrics proposed to describe the relation between the stimuli loudness spectrum and the MDS space, high correlations of $0.92 -$

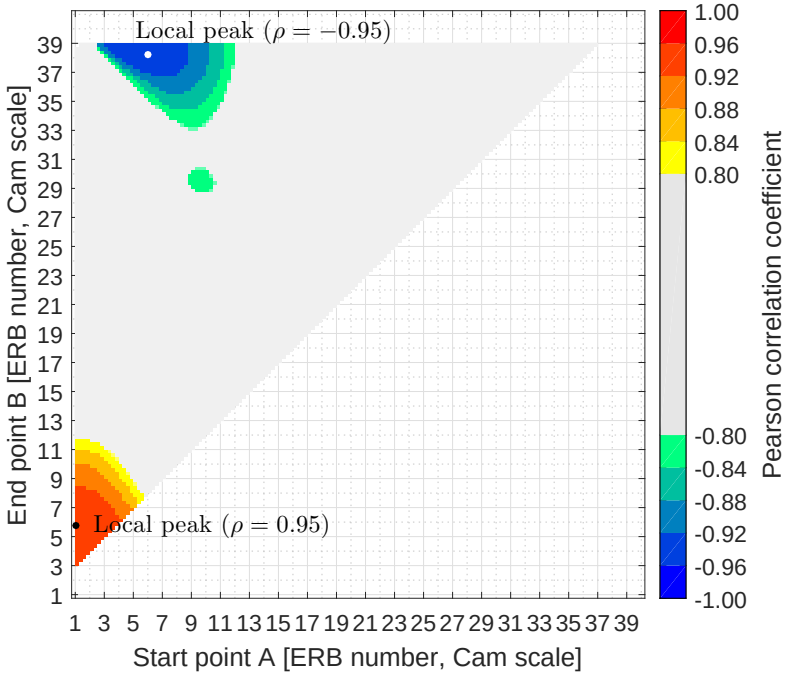


Fig. B.7: (Color online) Correlation between *DeepBass* and perceptual dimension 1 plotted as a function of the start frequency A and end frequency B used in the caculation of the metric based on 'Ch. average' data. The lightest grey represents all values within ± 0.8.

0.95 (averaged across excerpts) were found, showing that the metrics accurately describe the two perceptual dimensions for this dataset. The high correlations support the hypotheses that the two dominating MDS dimensions were both related to spectral characteristics. Processing the stimuli using a time-varying loudness model [27] was suited for describing the perceptual differences between headphones in this study. Since the selected musical excerpts had limited stereo effects, the headphones' differences in spatial sound qualities may not have been distinct. It is possible that excerpts having a more spatial quality would excite additional perceptual dimensions and even that these dimensions could account for a larger part of the perceived differences than the dimensions found in the present study. It is however a sign of general validity that two excerpts of such different genres led to the same main dimensions.

Both metrics related to perceptual dimension 1 were associated with the deep/low bass content ($20 - 222\,Hz$). Based on the present dataset it could not be concluded whether the perceptual dimension was related to deep bass relative to the full spectrum or relative to the narrower frequency range $246\,Hz - 3.2\,kHz$. *DeepBass* and *Bass/Mid* led to the same correlation level (0.95 averaged across excerpts) and the two metrics were highly correlated (0.94). For perceptual dimension 2, the two metrics *Bass+Mid* and *Mid/Treble* led to very similar correlations: 0.93 and 0.92 respectively and the two metrics were also highly correlated (0.79). It is therefore inconclusive whether the second perceptual dimension was related primarily to the relative loudness in the lower midrange or the ratio between the lower midrange and the treble range. The similar correlations of *Bass+Mid* and *Mid/Treble* may stem from whether or not the upper midrange was part of the listeners' decision criteria. In general, the uncertainty regarding the involvement of loudness in the midrange for both dimensions 1 and 2 could suggest that some listeners based their decisions on the full frequency range and others on more narrow frequency areas. A cluster analysis of listeners did not show any clearly separated clusters, but further work is needed to fully understand the decision process of listeners.

In [18], the sound reproductions of 37 single loudspeakers in a room were compared. Three main dimensions were highlighted as used by the listeners to differentiate these reproductions. While one dimension was clearly spatial and associated with the interaction of the loudspeakers with the room, the two other dimensions were spectral and very comparable to the dimensions obtained in the present study. These dimensions were also described using metrics defined on the specific loudness of the stimuli (recordings of a short musical excerpt reproduced by each loudspeaker in the room, reproduced using headphones during the listening test). These metrics, *BassTreb* and *Mid2*, are shown in Eq. B.5 and Eq. B.6 respectively. Note that [18] calculated the specific loudness spectrum, $Dens_m(f)$, using a different model [29] than

the one used in the present study.

$$BassTreb = \frac{\sum_{f=20\,Hz}^{280\,Hz} Dens_m(f)}{\sum_{f=1.8\,kHz}^{15.5\,kHz} Dens_m(f)} \tag{B.5}$$

$$Mid2 = \frac{\sum_{f=20\,Hz}^{280\,Hz} Dens_m(f) + \sum_{f=1.8\,kHz}^{15.5\,kHz} Dens_m(f)}{\sum_{f=280\,Hz}^{1.8\,kHz} Dens_m(f)} \tag{B.6}$$

Table B.3: Pearson correlation coefficients between MDS dimensions of the present study and metrics proposed in a loudspeaker study [18]. The two numbers in each cell represents the correlation for the Todd Terje and Tina Dickow excerpts respectively. Correlation with a * are significant on a $\alpha = 0.05$ level.

Metric	Dim.	Left ch.	Right ch.	Ch. average
BassTreb	1	0.73*/0.68*	0.84*/0.82*	0.82*/0.78*
Mid2	1	0.40/0.26	0.42*/0.19	0.43*/0.22
BassTreb	2	0.46*/0.67*	0.31/0.54*	0.39/0.60*
Mid2	2	-0.86*/-0.78*	-0.63*/-0.85*	-0.79*/-0.88*

BassTreb is the ratio between the specific loudness in the bass and upper midrange & treble. *Mid2* is the inverse ratio of the lower midrange and the specific loudness in the remaining spectrum. The two metrics, *BassTreb* and *Mid2*, were calculated for the stimuli of the current study in order to compare the dimensions obtained across studies. Table B.3 shows the correlation of *BassTreb* and *Mid2* with the perceptual dimensions 1 and 2. *BassTreb* is highly correlated with dimension 1 ($\rho = 0.80$), while *Mid2* is highly correlated with dimension 2 ($\rho = -0.84$). The difference in sign for *Mid2* has no significance as the MDS dimensions found using the metric MDS method are subject to inversion. The correlations between *BassTreb* and *Mid2* with the metrics of the current study are shown in Table B.4. The highest correlations with *BassTreb* are obtained with *DeepBass* and the highest correlations with *Mid2* are obtained with *Bass+Mid*. The levels of correlation reported in Table B.3 and B.4 suggest that the perceived dimensions could have been similar in the two studies despite the differences in reproduction systems and musical excerpts. More data would however be needed to establish which metrics from the current study corresponds best with BassTreb and *Mid2*.

In Gabrielsson and Sjögren, a questionnaire submitted to 40 sound engineers on the suitability of 200 adjectives (sensory descriptors) for evaluation of audio reproduction led to 30 adjectives being used to describe differences

Table B.4: Pearson correlation coefficients between the metrics *BassTreb* and *Mid2* proposed in a loudspeaker study [18] and the four metrics proposed in the current study. The metrics were calculated from the 'Ch. average' data. The two numbers in each cell represents the correlation for the Todd Terje and Tina Dickow excerpts respectively. Correlation with a * are significant on a $\alpha = 0.05$ level.

Metric	*DeepBass*	*Bass / Mid*	*Bass+Mid*	*Mid / Treble*
BassTreb	0.89*/0.79*	0.67*/0.59*	0.55*/0.80*	0.53*/0.57*
Mid2	0.37/0.29	0.66*/0.54*	-0.78*/-0.65*	-0.42/-0.32

in reproduction of headphones. Out of these 30 adjectives, only three were directly related to spectral characteristics (emphasized bass, bright/light, and full), although more may be strongly related (e.g. nasal, thin, dull, rumbling). An experiment was then conducted in which 20 subjects rated the quality of the 30 adjectives for eight headphones. A factor analysis resulted in five main factors of which the fourth was found clearly related to spectral characteristics: "brightness-darkness", with the strongest adjective correlation being to "emphasized bass". While this corresponds well to the findings reported in the present study of (deep) bass being a significant factor, the study of [7] showed less emphasis on spectral characteristics, with three factors explaining a higher proportion of the variance in the evaluations. These factors (sharpness, clearness and disturbing sounds) may result from headphones having a generally lower sound quality with clearly perceivable artefacts, which may not have been as distinct in the present headphones selection. The fifth factor, feeling of space, which was not found in the present study, may have been a consequence of the selected musical excerpts (four of the five musical excerpts in [7] were recorded in large rooms and included choirs or orchestras) or differences in spacial aspects caused by the wider selection of transducer technologies (electrodynamic, electrostatic, orthodynamic, and piezo-electric) used by Gabrielsson and Sjögren.

In [6], a higher emphasis on the spectral reproduction was reported in an analysis of ten trained listeners' comments on their reasoning behind preference ratings of six headphones. They investigated the correlation between the number of the occurrences of sensory descriptors and the preference scores. The highest correlated descriptor was good spectral balance (-0.92) and a total of 9 out of 19 descriptors were directly related to spectral characteristics (good bass extension, mid/treble peak, colored, etc.), with 5-6 others possibly indirectly related (e.g. dull, veiled, boomy, etc.). The fact that these descriptors are closely related to the dimensions highlighted in the present study might indicate that preference is related to the dominating perceptual differences between headphones. The second and third highest correlations were however distorted and wide sound stage, two aspects of the reproduction which were not highlighted in the present study, although informal debrief-

ing included a few listeners reporting that they took spatial characteristics into account in the evaluation of dissimilarity (they mentioned: externalisation, localization, envelopment, reverberation and room-size perception).

A dominance of spectral characteristics was also found in the study of [8], in which six headphones were evaluated. A panel discussion between trained listeners led to eight sensory descriptors needed for characterisation of active noise control headphone reproduction in background noise, among which four were related to spectral characteristics. A hierarchical multiple factorial analysis resulted in four main dimensions: timbre/attenuation, dynamic/spatial, precision/stereo space, and treble range; i.e. with the first and the fourth directly related to spectral characteristics. The spectral attributes selected in [8] included bass strength, treble strength and treble range (extension). While bass and treble are also elements in the four metrics proposed in the current study, treble range was not; in contrary the metric *DeepBass* maybe viewed as bass extension as opposed to treble extension. Like for [7] and [6], a dimension was found related to spatial characteristics, possibly a consequence of the selected musical excerpts as discussed previously.

The auralization process described in Section II B was designed to compensate for the spectral colouration of both the recording process and the reproduction of the Playback headphones. In Figure B.4 and Figure B.5 the results showed the unprocessed musical excerpt (Original) and the auralized Playback headphone to have overlapping confidence ellipses. The confidence ellipses (depicted only for these two stimuli for the sake of clarity) were estimated using a bootstrap method with 500 iterations. Furthermore, the playback headphones were, for both excepts, the headphones positioned closest to the Original. Ideally the perceived difference between the auralized headphones (Playback) and the real headphones (Original) should be smaller than the smallest difference measured in the other comparisons of headphones. Even though this was not the case in the current experiments, this difference seems close to the smallest perceived difference.

The debriefing following the listening tests showed that a few listeners reported overall loudness differences between stimuli. Three listeners in one experiment (Todd Terje) and two in the other experiment (Tina Dickow). None of the listeners participating in both experiments reported loudness differences in both. Considering the normal variation in hearing thresholds and high sensitivity concerning loudness differences (just noticeable differences of $0.5\,dB$ were found for typical audio programmes material in [30]), it was expected that not all listeners would perceive loudness as the listeners performing the informal loudness equalisation. Since none of the listeners, participating in both experiments, reported loudness as a type of difference perceived in both experiments suggests that overall loudness differences did not influence the results.

The most reported type of difference from the debriefing was left-right

balance differences between pairs. The left-right balance was purposefully preserved in the auralization process, to keep all headphones characteristics and thereby the differences between them. Left-right balance was however not found related to any of the MDS dimensions. A correlation analysis comparing the perceptual dimensions to the mean and max left-right differences in the loudness spectrum showed no significant correlations ($p > 0.2$).

5 Conclusions

A pair-wise comparison study was conducted where listeners were asked to evaluate the dissimilarity between 21 headphone reproductions of two musical excerpts. The stimuli consisted of the musical excerpts reproduced over the headphones, recorded, post-processed and reproduced over a set of playback headphones. The recorded headphones were of a wide range of types and prices. A multidimensional scaling analysis showed that two dimensions dominated the perceptual evaluation of dissimilarity. Furthermore it was shown that these dimensions could be accurately modelled by metrics based on the reproduced stimuli spectrum. A correlation analysis led to two competing metrics for each dimension. The first perceptual dimension was found to be related to either the bass loudness relative to the full-band loudness ($\rho = 0.95$) or relative to the loudness of the midrange frequencies exclusively ($\rho = 0.95$). The second dimension was found to be related either to the bass-midrange loudness relative to the full-band loudness ($\rho = 0.93$) or relative to the treble loudness exclusively ($\rho = 0.92$).

Acknowledgement

This work was funded by DELTA, ENTPE, and the Danish Agency for Science, Technology and Innovation (Case number: 1355-00061). Part of this work was performed within the Labex CeLyA (ANR-10-LABX-0060/ ANR-11-IDEX-0007). The authors wish to thank Torben H. Pedersen for good discussions on the methodology, Tore Stegenborg-Andersen for making many of the headphone recordings, Kevin Perreaut for assistance with loudness equalisation and pilot testing, as well as all the listeners who took part in the listening tests. Furthermore we wish to thank the three anonymous reviewers for helpful comments and suggestions.

References

[1] D. W. Martin and L. J. Anderson, "Headphone Measurements and Their Interpretation," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 63–70, 1947.

[2] G. Theile, "On the Standardization of the Frequency Response of High-Quality Studio Headphones," *J. Audio Eng. Soc*, vol. 34, no. 12, pp. 956–969, 1986.

[3] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design Criteria for Headphones," *J. Audio Eng. Soc*, vol. 43, no. 4, pp. 218–232, 1995.

[4] G. Lorho, "Subjective Evaluation of Headphone Target Frequency Responses," in *Proc. Audio Engineering Society Convention 126*, (Munich, Germany), Audio Engineering Society, May 2009.

[5] S. B. Chon and K.-M. Sung, "Sound Quality Assessment of Earphone: A Subjective Assessment Procedure and an Objective Prediction Model," in *Proc. Audio Engineering Society Conference: 38: Sound Quality Evaluation*, (Piteå, Sweden), pp. 1–8, Audio Engineering Society, June 2010.

[6] S. Olive and T. Welti, "The Relationship between Perception and Measurement of Headphone Sound Quality," in *Audio Engineering Society Convention 133*, Oct. 2012. Convention Paper 8744.

[7] A. Gabrielsson and H. Sjögren, "Perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 1019–1033, 1979.

[8] N. Zacharov, J. Ramsgaard, G. Le Ray, and C. V. Jørgensen, "The multidimensional characterization of active noise cancelation headphone perception," in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, (Trondheim, Norway), pp. 130–135, IEEE, June 2010.

[9] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3678–3688, 1996.

[10] G. Lorho, *Perceived Quality Evaluation, An Application to Sound Reproduction over Headphones*. Ph.D. thesis, Department of Signal Processing and Acoustics, School of Science and Technology, Aalto University, Espoo, Finland, 2010. ISBN 978-952-60-3196-5 (Electronic).

[11] M. Opitz, "Headphones Listening Tests," in *Audio Engineering Society Convention 121*, (San Francisco, CA, USA), Audio Engineering Society, Oct. 2006. Convention Paper 6890.

[12] A. Silzle, B. Neugebauer, S. George, and J. Plogsties, "Binaural Processing Algorithms: Importance of Clustering Analysis for Preference Tests," in *Audio Engineering Society Convention 126*, Audio Engineering Society, 2009. Convention paper 7728.

[13] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences," *J. Acoust. Soc. Am.*, vol. 117, no. 6, pp. 3832–3840, 2005.

[14] J. M. Murray and C. M. Delahunty, "Mapping consumer preference for the sensory and packaging attributes of Cheddar cheese," *Food Qual. Prefer.*, vol. 11, no. 5, pp. 419 – 435, 2000.

[15] T. H. Pedersen and N. Zacharov, "The Development of a Sound Wheel for Reproduced Sound," in *Audio Engineering Society Convention 138*, (Warsaw, Poland), pp. 1–13, Audio Engineering Society, May 2015. Convention Paper 9310.

[16] M. Lavandier, P. Herzog, and S. Meunier, "Comparative measurements of loudspeakers in a listening situation," *J. Acoust. Soc. Am.*, vol. 123, no. 1, pp. 77–87, 2008.

[17] M. Lavandier, S. Meunier, and P. Herzog, "Identification of some perceptual dimensions underlying loudspeaker dissimilarities," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4186–4198, 2008.

[18] P.-Y. Michaud, M. Lavandier, S. Meunier, and P. Herzog, "Objective Characterization of Perceptual Dimensions Underlying the Sound Reproduction of 37 Single Loudspeakers in a Room," *Acta Acust United Ac*, vol. 101, pp. 603–615, May 2015.

[19] M. Paquier and V. Koehl, "Discriminability of the placement of supra-aural and circumaural headphones," *Appl. Acoust.*, vol. 93, pp. 130–139, June 2015.

[20] F. Brinkmann, A. Lindau, and Z. Schaerer, "Regulated Inversion, a MATLAB script," Mar. 2015.

[21] S. G. Norcross, M. Bouchard, and G. A. Soulodre, "Inverse Filtering Design Using a Minimal-Phase Target Function from Regularization," in *Audio Engineering Society Convention 121*, (San Francisco, USA), Audio Engineering Society, Oct. 2006. Convention paper 6929.

[22] N. Cowan, "On short and long auditory stores," *Psychological Bulletin*, vol. 96, no. 2, pp. 341–370, 1984.

[23] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. Springer series in statistics, New York, NY: Springer, 1. ed ed., 1997.

[24] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[25] C. P. Volk, S. Bech, T. H. Pedersen, and F. Christensen, "Five Aspects of Maximizing Objectivity from Perceptual Evaluations of Loudspeakers: A Literature Study," in *Audio Engineering Society Convention 138*, (Warsaw, Poland), pp. 1–12, Audio Engineering Society, May 2015. Convention Paper 9230.

[26] J. Kruskal and M. Wish, *Multidimensional scaling*. California, USA: Sage Publications, 1978.

[27] B. R. Glasberg and B. C. J. Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *J. Audio Eng. Soc*, vol. 50, no. 5, pp. 331–342, 2002.

[28] H. Hudde, A. Engel, and A. Lodwig, "Methods for estimating the sound pressure at the eardrum," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1977–1992, 1999.

[29] E. Zwicker and H. Fastl, "A portable loudness-meter based on ISO 532 B," in *Proc. 11th International Congress on Acoustics*, (Paris, France), pp. 135–137, 1983.

[30] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "The Subjective Loudness of Typical Program Material," in *Audio Engineering Society Convention 115*, (New York, NY, USA), Audio Engineering Society, Oct. 2003. Convention Paper 5892.

Paper C

Modelling perceptual characteristics of prototype headphones

Christer P. Volk, Torben H. Pedersen,
Søren Bech, and Flemming Christensen

The paper has been published as Open Access in the
Proc. of Audio Engineering Society International Conference on Headphone Technology,
pp. 1–-9, Aalborg, Denmark: Audio Engineering Society, 2016.

To accommodate the smaller format of the printed thesis publication, the layout has been revised. The text is identical to the submitted/published article.

Abstract

This study tested a framework for modelling of sensory descriptors (words) differentiating headphones. Six descriptors were included in a listening test with recordings of the sound reproductions of seven prototype headphones. A comprehensive data quality analysis investigated both the performance of the listeners and the suitability of the descriptors for modelling. Additionally, two strategies were investigated for modelling metrics describing these descriptors, both relying on specific loudness estimations of the test stimuli. The stability of the initially found metrics was tested with a bootstrap procedure to quantify the potential of the metrics for future predictions within the perceptual space spanned by the headphones. The most promising results were metrics for Bass, Clean and Dark-Bright with correlations values of $r^2 = 0.62$, $r^2 = 0.58$, *and* $r^2 = 0.90$ *respectively.*

1 Introduction

The development of headphones can be a process involving many prototypes while exploring the potential of new technologies, constructions, or materials. At some point in the process, the many paths explored must be narrowed down to a single track. This study explores the possibility of modelling the link between the physical sound reproduction and the perceptual characteristics of headphones, thereby potentially providing an indication of which path will lead to the desired design target.

Perceptual characterisation of sound reproduction offers evaluation of performance based on human perception as an alternative to traditional electro-acoustical measurements, such as frequency- and time responses. It can provide valuable insight into the dominating perceptual characteristics of e.g. a set of headphones. An issue with perceptual evaluations is however the time and resources required to collect data. Consequently, a number of mathematical models based on less resource-heavy electro-acoustic measurements have previously been proposed. They were able to predict preference [1], mean opinion score [2, 3] or stereo width [4], thereby providing insight as to sound quality performance. Earlier efforts with prediction of loudspeaker preferences from visual inspection of frequency responses were also attempted by Toole in [5], who concluded that: *"Listeners, it seems, like the sound of loudspekaers with a flat, smooth wideband on-axis amplitude response that is maintained at substantial angles off axis"*.

Within external preference mapping [6], the underlying decision process leading to a listener's preference is assumed to be a weighted sum of perceived auditory characteristics. These weights are based on a personal reference consisting of desired features and can be influenced by prior experience, context, mood etc. [7]. The concept can be described by Eq. C.1 for a product

i. $S_X(i)$ represents the salience of the characteristics described by a sensory descriptor X and the constants α to ω are an individual's weightings of the characteristics' importance for preference. ϵ denotes the residual. Note that the relationship between S_X terms may be non-linear, although a linear case is illustrated in Eq. C.1.

$$Preference(i) = \alpha S_1(i) + \beta S_2(i) + \cdots + \omega S_N(i) + \epsilon \tag{C.1}$$

The weights are individual and subjective, while the salience of an auditory characteristic, S_X, is considered objective and depending only on the auditory acuity of listeners. Leaving out the subjective weights, not all S_X-terms are relevant for product characterisation of a given subset of audio reproduction products being evaluated (compared). A limited number of terms are likely to dominate the overall sensation, but which terms that is will depend on the products being evaluated. Differences in dominating characteristics are for instance seen between the headphones study [8] by Gabrielsson and Sjögren from 1979 and the headphones study [9] by Olive and Welti from 2012. In the older study, the analysis led to characteristics with emphasis on artefacts, while the newer study led to characteristics with emphasis on spectral differences.

In the present study a framework was established and tested for finding metrics able to predict the perceptual characteristics of headphones sound reproductions. Recordings were made of seven prototype headphones' reproduction of a selection of musical excerpts. The recordings were used as stimuli in a listening test as well as the basis on which proposed metrics were calculated. The metrics were thereby developed directly on the basis of what listeners perceived. This approach was also used in e.g. [10] to study the dominating perceptual dimensions differentiating monophonic loudspeaker reproduction in a room. The listening test of the present study consisted of evaluation of a number of sensory descriptors[1] by expert listeners, with the purpose of characterising the perceptual space spanned by the headphones. The perceptual ratings were modelled on the basis of estimated stimuli loudness spectra, and consequently the non-linearities of the human auditory processing are incorporated in the proposed metrics.

2 Listening test

2.1 Headphones

A total of eight electrodynamic headphones were included in this study. Seven were prototype models from one manufacturer, and one additional

[1] A sensory descriptor is defined here as a word or phrase that describes, identifies, or labels a perceptual characteristic of a system, e.g. a headphone reproduction. This definition is adapted from [11].

set of high-end headphones was included as a reference. The seven prototypes consisted of four supraaural and three circumaural. All the prototypes were closed-back headphones, while the reference headphone model was circumaural and open-back.

2.2 Stimuli

Four musical excerpts were played over the eight headphones. The reproduced audio was recorded, post-processed, and presented to listeners over a pair of Sennheiser HD 650 headphones (playback headphones). In the recording process the headphones were placed on a Brüel & Kjær 4128C head- and torso simulator (B&K HATS). To minimize (asymmetrical) leakage, the headphone positioning was checked by recording pink noise and comparing the right-left input level balance. The amplifier gain was adjusted for each set of headphones to record at approximately the same sound level across all the headphones (mean $L_{eq} = 69.9\,dB\,(A)$, $\sigma = 5.4\,dB$, variation dominantly due to musical excerpt differences). The binaural recordings were captured using a RME Fireface 800 soundcard with a 24 bit A/D converter and saved at sample rate of $48\,kHz$.

The recordings were post-processed to compensate for the influence of the ear-canals of the B&K HATS as well as for the frequency response of the playback headphones. The compensation was performed by means of a equalizer with 1/3-octave band minimum-phase FIR filters on both channels, i.e. without compensation for left-right imbalance in sensitivity. The filter had a dip in the range $80 - 400\,Hz$ with a minimum of $-1.4\,dB$ at $125\,Hz$ and another in the range $0.5 - 12.5\,kHz$ with a minimum of $-12\,dB$ at $3.15\,kHz$. The post-processed recordings were loudness normalized using an automated process, where loudness was estimated using a stationary loudness model [12] and iteratively level-adjusted to reach a target (channel-averaged) of equal loudness at a playback level of $67 \pm 0.01\,Phon$. Finally, the stimuli were converted to 16 bit WAV files for reasons of compatibility with the test software. The post-processed headphone recordings are referred to as auralised headphones in the following.

A total of ten musical excerpts were originally recorded. During a session of informal listening by two experienced listeners, four were selected for the listening test, as they were perceived to facilitate the largest discrimination between headphones: Jennifer Warnes ('Bird on a Wire', Famous Blue Raincoat, 1987), Todd Terje ('Delorean Dynamite', It's album time, 2014), Helge Lien Trio ('Natsukashii', Natsukashii, 2011), and George Druschetzky Ensemble Zefiro ('Serenata for winds & strings in E flat major: Maestoso, Allegro', Druschetzky: Quartetto; Serenata; Quintetto, 2002). These four excerpts represent the musical genres: Pop, Electronic, Jazz, and Classical respectively.

2.3 Test procedure

The listening test comprised of evaluations with six sensory descriptors selected as suited for discriminating between the headphones. The selection was based on consensus meetings with trained listeners [13], specifically trained in the sensory descriptors described in the DELTA-developed Sound wheel [11] and all descriptors were consequently selected among these. The chosen descriptors were: Bass strength, Midrange strength, Treble strength, Dark-bright, Clean, and Punch. They comprised sensory descriptors from three main groups: Dynamics, Timbre and Transparency. Danish names and definitions were used, as all listeners were native Danish speakers.

The listening test consisted of evaluation of the eight headphones with regards to each sensory descriptor on a $15\,cm$ rating scale anchored by two words specific for each descriptor. The reference headphones were included both as a labelled reference and as a hidden anchor system. The definitions included instructions on which rating to give the hidden anchor (if identified). Consequently the ratings of the reference headphones were excluded from the data analysis. The SenseLabOnline test software (senselabonline.com), allowed listeners to listen to each stimulus as many times as needed and switch between stimuli almost instantaneously. One "screen" in the user interface included stimuli for one musical excerpt and evaluation on one sensory descriptor with all auralized headphones. The full test comprised 48 "screens" (six descriptors, four musical excerpts and 2 repetitions) presented in randomised order and evaluated during one 2-hour session per listener. Listeners were encouraged to take breaks on a regular basis. The playback level during the test was set to approximately $80\,dB(A)$ (measured with the playback headphones positioned on a B&K HATS at the default level setting). The listeners did however have the option to ask the test leader for small level adjustments ($\pm 4\,dB$) during familiarisation to accommodate a comfortable listening level for the individual. Level adjustments affect the perception of the spectral balance of the stimuli, but makes the long sessions more comfortable for listeners and have a small influence in comparison to the natural variation in hearing between listeners.

2.4 Listeners

Eighteen listeners participated in the listening test. All were trained listeners from DELTA SenseLab's expert panel. Among the 18 listeners eight were trained specifically in the sensory descriptors of the Sound wheel [11]. The listeners ranged in age from 20 to 54 with a median of 29, and all had their hearing tested both prior to joining the expert panel as well as periodically afterwards.

2.5 Listening test results

Listener performance

The performance of the participating listeners was evaluated per sensory descriptor by means of two statistical measures, which will be briefly explained: the eGauge metrics Discrimination and Reproducibility [14] as well as Tucker-1 analysis [15]. Discrimination describes a listener's ability to statistically discriminate between systems (headphones), i.e. a measure of how big an influence the systems have on the ratings - as opposed to other factors, such as the influence of musical excerpts, conditions, etc. Reproducibility describes how consistent a listener rate the same stimuli (a headphone and musical excerpt combination in this case) between repetitions. The Tucker-1 analysis is based on a Principal Component Analysis (PCA) and was used here to gauge listener performance in terms of consistency and agreement with the panel average/consensus. Listeners below the noise floor (performance

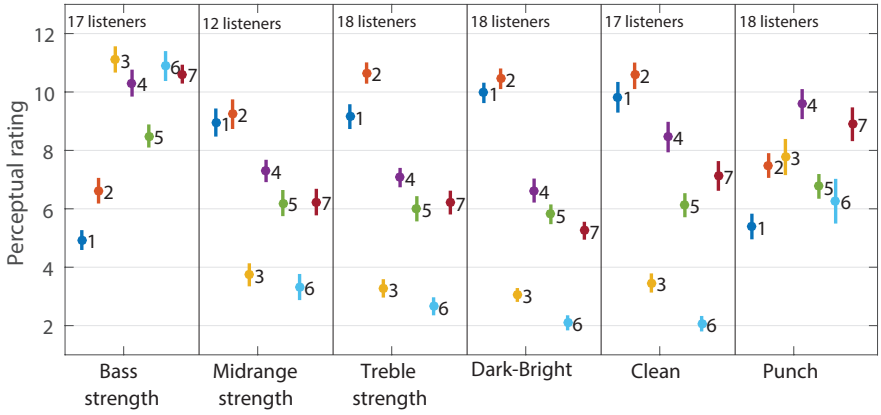


Fig. C.1: [Colors online] Mean and 95% confidence intervals of the six sensory descriptor ratings. Each point and number represents the rating of one auralized headphone averaged over selected listeners, musical excerpts, and repetitions. The number of listeners included for each descriptor is displayed above the ratings.

of random evaluations), with regards to Discrimination and Reproducibility, were removed from the dataset prior to further analysis, as was listeners within the inner ring (less than 50% explained variance) of the Tucker-1 loading scores plot. This meant removing six listeners from Midrange strength, and one listener from both Bass strength and Clean. Midrange strength was thus a difficult sensory descriptor to evaluate for the listeners on the presented stimuli. In addition, the Tucker-1 analysis of the sensory descriptors, showed a wide spread of listeners' ratings within the two circles for the Punch sensory descriptor, which signifies lack of agreement between listeners. This

Variable	MS	F	Pr > F
Sys (**Bass str.**)	79000	356	< 0.0001
Sys/sample	1630	7.33	< 0.0001
Sys (**Midrange str.**)	51300	233	< 0.0001
Sys/sample	187	1.58	0.06
Sys (**Treble str.**)	120000	626	< 0.0001
Sys/sample	1220	6.36	< 0.0001
Sys (**Dark-Bright**)	145000	701	< 0.0001
Sys/sample	1950	9.45	< 0.0001
Sys (**Clean**)	137000	448	< 0.0001
Sys/sample	507	1.67	0.04
Sys (**Punch**)	30800	82.3	< 0.0001
Sys/sample	2880	7.68	< 0.0001

Table C.1: System effect (6 DF) and system/sample interaction effect (18 DF) from 4-ways ANOVA tables for each sensory descriptor. Sys/sample is the interaction between headphone and musical excerpt. Mean square (MS) and F-values are rounded to three significant digits.

could be caused by e.g. listeners' needing more training or sensory descriptors which are not well-defined or ill-suited for the purpose.

Sensory descriptor assessment

In Fig. C.1 the mean ratings for each sensory descriptor are depicted[2], based on ratings of the listeners that was not removed as a result of the performance criteria described in the previous section.

In terms of modelling perspectives, the most important prerequisite, was considered to be the sensory descriptors ability to discriminate between the headphones, e.g. to model Bass strength, the ratings of the headphones was required to be significantly different from each other and preferably span the majority of the rating scale. A 4-ways (Headphone $\times$ Musical excerpt $\times$ Listener $\times$ Repetition) fixed-effect analysis of variances (ANOVA) was conducted for each descriptor to test its ability to discriminate between headphones. The results are presented in Table C.1 and showed that all sensory descriptors were able to discriminate between two or more of the headphones. Punch, however, had the lowest F-value (discriminatory power) and combined with poor agreement in the Tucker-1 analysis, no efforts were done to model this descriptor. The interaction between headphones and musical excerpt was significant for all descriptors with the exception of Midrange strength ($p \leq 0.05$). A similar p-value was however seen for Clean as well.

[2]The interaction between factors 'Headphone' and 'Musical excerpt' is significant for most sensory descriptors (see Table C.1), but its influence (F-value) is at least one order of magnitude

r	Bass	Dark-Bright	Mid	Treble	Clean	Punch
Bass	**1.00**	-0.87	-0.81	-0.79	-0.72	0.58
Dark-Bright	-0.87	**1.00**	**0.97**	**0.97**	**0.94**	-0.17
Mid	-0.81	**0.97**	**1.00**	**0.97**	**0.98**	-0.05
Treble	-0.79	**0.97**	**0.97**	**1.00**	**0.96**	-0.03
Clean	-0.72	**0.94**	**0.98**	**0.96**	**1.00**	0.08
Punch	0.58	-0.17	-0.05	-0.03	0.08	**1.00**

Table C.2: [Colors online] Pearson correlation coefficients for the sensory descriptor ratings. The descriptors are sorted in descending order of their absolute correlation with bass. The intensity of the background color represent the degree of correlation. Bold numbers have $r > |0.90|$.

A correlation analysis of the sensory descriptor ratings, based on mean values from well-performing listeners, are shown in Table C.2. The correlations seen to be high between all descriptors - with the exception of Punch. Due to the listeners' disagreement in the rating of Punch it was however not possible to know whether the descriptor comprised an important perceptual dimension or simply a dimension of noise. It is noteworthy that Dark-Bright had a higher correlation with Treble strength than with Bass strength. Furthermore Clean was highly correlated with Midrange strength, Treble strength as well as Dark-Bright. While the ratings of four sensory descriptors were highly correlated it remained of interest to model all as each potentially represented a separate coupling to the physical world. Consequently they had varying performance potentials, making it of interest to model all and select the most promising in terms of prediction capabilities. Note, however, that correlations between sensory descriptors, does not imply that the sensory descriptors refers to the same percept in general, as similarity between this subset of headphones would lead to high correlations as well.

3 Modelling methodology

Metrics were developed to investigate the link between the sensory descriptor ratings and the listening test stimuli. They were all based on a stationary loudness model [12], where specific loudness was estimated for each stimulus in steps of 0.1 Bark from Bark number 0.1 (20 Hz) to 24.0 (15.5 kHz). The resulting loudness metrics (presented later on in Table C.3) are the summed loudness in a frequency range AB divided by the sum of the full loudness spectrum, as described by Eq. C.2. $Dens_m(f)$ is the temporal mean of the time-varying specific loudness, while A and B denotes the frequency limits

lower than the main 'Headphone' factor. Consequently, Fig. C.1 shows the average over excerpts.

of the AB range.

$$metric = \frac{\text{AB range}}{\text{Full range}} = \frac{\sum_{f=A}^{B} Dens_m(f)}{\sum Dens_m} \tag{C.2}$$

An optimisation routine was implemented to find the frequency range AB, at which the metrics had maximum correlation with the sensory descriptor ratings. As it was not of interest to get metrics related specifically to individual musical excerpts, the routine found the frequency range where the maximum average correlation across excerpts were located, as described by Eq. C.3 and C.4. *Metric* is a matrix with $Dens_m(AB)$ loudnesses for all tested combinations of AB frequency ranges. P_{ex} is the average perceptual ratings for each set of headphones for the musical excerpt ex. R_{ex} is the Pearson correlation matrix for all tested combinations of AB frequency ranges. R_{max} is the maximum Pearson correlation averaged over all P_{ex}. A metric's AB frequency range was consequently the range where R_{max} was found.

$$R_{ex} = corr\left(Metric, P_{ex}\right) \tag{C.3}$$

$$R_{max} = max\left(\frac{\sum_{ex=1}^{4} R_{ex}}{4}\right) \tag{C.4}$$

All combinations of search ranges in the full loudness spectrum and all search positions were tested with the 0.1 Bark resolution with one constraint: The search range was restricted to AB ranges with a minimum width of two Barks, to avoid getting peak correlations in narrow ranges unlikely to be the cause of the perceptual rating. In the remaining part of this paper, the output of the optimisation routine, when analysed with all seven prototype headphones, are referred to as the 'baseline'.

The output of the optimisation routine was likely to lead to several peaks due to the wide search area. This ensured a search unbiased by the authors' theories, but complicated the analysis. Even if one peak had a (significantly) higher correlation coefficient, this may not have been the case with slightly different prototypes. This issue is commonly dealt with in the literature by training the metric on one set of data and validating the results with a separate dataset. This is however not desired with small datasets, such as a limited number of prototype headphones. Therefore, a bootstrap method was used to get a better representation of the sample space spanned by all prototype headphones. A total of 500 bootstrap iterations (sampling with replacement) of the optimisation routine were processed for each descriptor, followed by a classification task: For each iteration the optimisation routine output's maximum peak was classified either as matching one of the peaks

from the baseline or an unclassified peak ("Other"). A match was defined as: A and B from the AB range being within $\pm 2\,Bark$ of a peak from the baseline. Due to the limited number of headphones, this process was likely not to have a high hit rate, i.e. maxima's coinciding with the baseline peaks, but still allowed comparison of hit rates across peaks as well as providing estimated correlation coefficient confidence intervals.

3.1 Dark-Bright metric

For modelling of the Dark-Bright descriptor another approach was chosen. In the literature a metric commonly referred to as the spectral centroid (see e.g. [16, 17]) is reported to be a good predictor of brightness (a sensory descriptor similar to Dark-Bright). It is the balancing point in a spectrum, where an equal amount of energy is located below and above the point. In contrast to the cited papers the spectral centroid was, in the present study, calculated on the basis of loudness spectra rather than frequency spectra, as it was hypothesised to be a better predictor, due to the closer relation with the perception of listeners. The proposed Dark-Bright metric was therefore calculated as described by Eq. C.5. Since the output of the loudness model was in discrete 0.1 Bark bins, the solution became a minimization problem. $Dens_m(b)$ is the temporal mean of the time-varying specific loudness and b is the 0.1 Bark bin number. b_{MIN}, b_{CEN}, b_{MAX} are the minimum, centroid, and maximum bin numbers respectively. b_{CEN} thereby represents the point of equal loudness, i.e. the perceptual spectral centroid.

$$\begin{aligned} \min_{b_{CEN}\in\mathbb{Z}} \quad & \Big|\sum_{b=b_{MIN}}^{b_{CEN}} Dens_m(b) - \sum_{b=b_{CEN}+1}^{b_{MAX}} Dens_m(b)\Big| \\ \text{s.t.} \quad & \\ & b_{MIN} \geq b_{CEN} \leq b_{MAX} \end{aligned} \tag{C.5}$$

3.2 Metrics results

Numeric results of the optimization routine and bootstrap classification are shown in Table C.3. The routine output's a map showing correlation values for all processed AB ranges. The first column shows the multiple (competing) peaks for each sensory descriptor, e.g. three for Bass strength. The **Bass strength** P1 and P2 AB ranges pointed to the same conclusion: that the low-frequency range up to $210\,Hz$ was important for the perception of bass strength, i.e. P1 and P2 could be considered as equivalent. The 'Peak R' column displays the Pearson correlation coefficient peaks from the baseline with the perceptual data. The third peak, P3, was related to treble, implying that the level of high-frequencies may affect the perception of bass strength. The 'Hit rate' column shows that 26.2% of the bootstrap iterations led to P1

or P2 having the best correlation with the sensory descriptor Bass strength, with a median correlation coefficient for P1 of $r = -0.79$ and 95% CI's of $[-0.57$ to $-0.99]$. For **Midrange strength**, the peak with the highest correlation had an AB range within bass and low-midrange frequencies (again P2 was equivalent), while P4 was the only peak with a maxima coinciding with the baseline peaks in 500 of the bootstrap iterations. The correlation CI's of these bootstrap maxima was however inconsistent and spanned both positive and negative values, implying instability. For **Treble strength**, P2-P4 all led to high hit rates, with P2 having the highest hit rate and a narrow CI. P5 had a low hit rate, but covered the area traditionally considered the treble range. For **Clean**, P4 and P5 resulted in a combined hit rate of 31.6%. The bootstrap CI's for P4 spanned the smallest range of values. For the Dark-Bright sensory descriptor a bootstrap method was again employed to test the stability of the proposed metric. Pearson correlation coefficients for 500 bootstrap iterations had a median value of $r = 0.95$ and 95% CI's of $[0.81; 0.99]$.

The relation between the proposed metrics and the perceptual ratings are shown in Fig. C.2.

4 Discussion

For the sensory descriptors modelled using Eq. C.2, the output of the optimization routine led to multiple peaks with correlation coefficients $r^2 \geq 0.67$ ($r \geq |0.82|$), when all headphones were analysed. This approach was deemed appropriate for modelling of the four sensory descriptors presented in Table C.3. In the case of Bass-, Midrange- and Treble strength, they all had a peak, which logically seemed probable: An AB-range of $20 - 200\,Hz$ (P2) for Bass strength, an AB-range of $690 - 5900\,Hz$ (P3) for Midrange strength, and an AB range of $8.7 - 15\,kHz$ (P5) for Treble strength. The bootstrap categorisation method showed the peak P1 (and the equivalent P2) to have a promising hit rate (26.2%), while Midrange strength, in contrast to logic, showed the highest hit rate for an AB-range in the high-frequency region. For Treble strength three peaks got high hit rates ($20.0 - 29.6$%), none of which seem logically related to the perception of treble. In the case of Midrange- and Treble strength the bootstrapping process uncovered uncertainties in the data, but did not point to peaks likely to have a causal relation with perception. For Bass strength, the method showed that peak P1 may be a better predictor of bass strength, than the equivalent but more logical choice peak P2, due to numerical stability. For Clean peak P4 and the equivalent P5 got a combined hit rate of 31.6%, with P4 being more stable with regards to CI's. Here, the bootstrapping process revealed peak P4 as a potential better predictor of the descriptor Clean, than P1, although P1 had the largest r-value in the output from the optimization routine when all headphones were analysed (baseline).

Metric		Peak R	AB Range	Hit rate	Bootstrap R	95% CI's
Bass strength	P1	-0.97	210-15000	25.8 %	-0.79	-0.99; -0.57
	P2	0.96	20-210	0.4 %	0	
	P3	-0.82	8900-14000	11.2 %	-0.82	-0.90; -0.74
Midrange strength	P1	-0.98	20-610	0 %		
	P2	0.98	640-15000	0 %		
	P3	0.97	690-5900	0 %		
	P4	0.91	8900-15000	6.2 %	0.80	**-0.99; 0.89**
Treble strength	P1	-0.95	20-650	2.8 %	-0.97	-1.00; -0.95
	P2	0.95	680-15000	29.6 %	0.97	0.95; 0.99
	P3	0.95	730-5700	26.6 %	0.96	0.93; 1.00
	P4	0.93	2500-4500	20.0 %	0.98	0.95; 0.99
	P5	0.93	8700-15000	2.8 %	0.99	0.94; 1.00
Clean	P1	0.99	800-4800	0 %		
	P2	-0.98	20-720	0.8 %	-0.99	-1.00; -0.98
	P3	0.98	730-15000	0 %		
	P4	-0.90	20-8200	24.6 %	0.76	0.58; 0.97
	P5	0.90	8200-15000	7.0 %	0.77	**-1.00; 0.92**

Table C.3: Potential metrics describing sensory descriptors. The first column displays a metric, e.g. bass strength, and peaks (P1, P2, etc.) in the output map of the optimization routine (baseline). The two next columns display the Pearson correlation coefficient value of the baseline peaks and their AB ranges. The 'Hit rate' column displays percentage bootstrap iterations having maximum at the baseline peak. The last two columns display the median correlation of the iterations with a match as well as their 95% confidence intervals (CI) calculated from bootstrap percentiles. CI's in **bold** have a range spanning both positive and negative *r* values. All numbers are rounded to two significant digits.

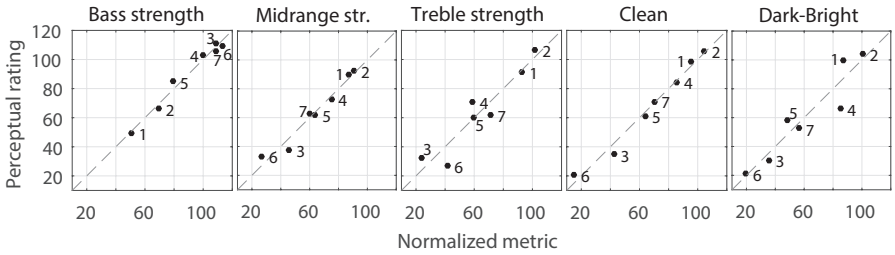


Fig. C.2: Scatterplot of perceptual ratings vs. fitted metric values. The metric values where normalised to the scale of the perceptual ratings (0-150). The metric values are based on P1 for Bass strength, P3 for Midrange strength, P5 for Treble strength, and P4 for Clean respectively.

In general, the many equivalent peaks may points to a problem in structure of the metrics described by Eq. C.2. For prediction of bass strength, a better metric could be $^{AB}/_{CD}$, i.e. with the denominator covering a limited frequency range CD rather than the full range. This approach was investigated in another study [18].

For Dark-bright the loudness spectral centroid correlated well with the perceptual data (median $r = 0.95$). Closer inspection of the relation between the metric's output and the perceptual ratings showed the auralized headphones '4' as a slight outlier for two of the four musical excerpts. Compared with the other headphones, these had an increased midrange within the frequency range $\approx 700 - 1400\,Hz$. This may indicate that wide resonances in the sound reproduction affect the perception of Dark-Bright slightly different than predicted by the proposed metric.

The three most promising metrics Bass strength, Dark-Bright and Clean models the sensory descriptors with the least correlation between them, thereby constituting a strong set of metrics for characterisation of the perceptual space spanned by the evaluated prototype headphones.

5 Summary

This paper presented a framework for modelling of sensory descriptors related to timbre of seven prototype headphones. Three metrics deemed stable was proposed for Bass strength ($r^2 = 0.62$), Clean ($r^2 = 0.58$), and Dark-Bright ($r^2 = 0.90$) respectively. All of them were based on loudness estimates of listening test stimuli. The first two were modelled from a simple equation (Eq. C.2) with specific loudness in an AB frequency range (found by optimisation) divided by loudness in the full spectrum. The metric proposed for Dark-Bright prediction was based on a spectral centroid calculation of specific loudness. For two other sensory descriptors, Midrange- and Treble strength, stability investigations based on a bootstrapping process revealed inconsistencies in the sign of the correlations or multiple competing local maxima.

Acknowledgement

This work was funded by DELTA and the Danish Agency for Science, Technology and Innovation (Case number: 1355-00061). The author wishes to thank the audio design company for supplying the prototype headphones, Tore Stegenborg-Andersen for processing of the stimuli and for conducting the listening test, as well as the two anonymous reviewers for valuable comments and suggestions.

References

[1] S. E. Olive, "A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model," in *AES Convention 117*, 2004. Convention paper 6190.

[2] ITU-R, "Method for objective measurements of perceived audio quality," Recommendation ITU-R BS.1387-1, International Telecommunication Union Radiocommunication Assembly (ITU-R), United States, 1998.

[3] ITU-T, "Perceptual objective listening quality assessment," Recommendation ITU-T P.863, ITU Telecommunication Standardization Sector (ITU-T), United States, Jan. 2011.

[4] M. Takanen and G. Lorho, "A Binaural Auditory Model for the Evaluation of Reproduced Stereophonic Sound," in *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*, pp. 1–10, Mar. 2012.

[5] F. E. Toole, "Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2," *J. Audio Eng. Soc*, vol. 34, no. 5, pp. 323–348, 1986.

[6] J. A. McEwan, "Preference Mapping for Product Optimization," in *Multivariate Analysis of Data in Sensory Science*, vol. 16 of *Data Handling in Science and Technology*, pp. 71–102, Elsevier, 1st ed., 1996.

[7] J. Blauert and U. Jekosch, "A Layer Model of Sound Quality," *J. Audio Eng. Soc*, vol. 60, no. 1/2, pp. 4–12, 2012.

[8] A. Gabrielsson and H. Sjögren, "Perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 1019–1033, 1979.

[9] S. Olive and T. Welti, "The Relationship between Perception and Measurement of Headphone Sound Quality," in *Audio Engineering Society Convention 133*, Oct. 2012. Convention paper 8744.

[10] M. Lavandier, P. Herzog, and S. Meunier, "Comparative measurements of loudspeakers in a listening situation," *J. Acoust. Soc. Am.*, vol. 123, no. 1, pp. 77–87, 2008.

[11] T. H. Pedersen and N. Zacharov, "The Development of a Sound Wheel for Reproduced Sound," in *Audio Engineering Society Convention 138*, (Warsaw, Poland), pp. 1–13, Audio Engineering Society, May 2015. Convention paper 9310.

[12] E. Zwicker and B. Scharf, "A model of loudness summation.," *Psychological Review*, vol. 72, no. 1, pp. 3–26, 1965.

[13] H. Stone, J. Sidel, S. Oliver, A. Woolsey, and R. C. Singleton, "Sensory Evaluation by Quantitative Descriptive Analysis," *Food Technology*, vol. 28, pp. 24–34, 1974.

[14] G. Lorho, G. Le Ray, and N. Zacharov, "eGauge—A Measure of Assessor Expertise in Audio Quality Evaluations," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, pp. 1–10, June 2010.

[15] P. M. Kroonenberg and J. de Leeuw, "Principal component analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, vol. 45, pp. 69–97, Mar. 1980.

[16] E. Schubert and J. Wolfe, "Does Timbral Brightness Scale with Frequency and Spectral Centroid?," *Acta Acustica united with Acustica*, vol. 92, pp. 820–825, 2006.

[17] I. B. Labuschagne and J. J. Hanekom, "Preparation of stimuli for timbre perception studies," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2256–2267, 2013.

[18] C. P. Volk, M. Lavandier, S. Bech, and F. Christensen, "Identifying the dominating perceptual differences in headphone reproduction," *Submitted, J. Acoust. Soc. Am.*, Feb. 2016.

Paper D

Modelling perceptual characteristics of loudspeaker reproduction in a stereo setup

Christer P. Volk, Søren Bech,
Torben H. Pedersen, and Flemming Christensen

The article has been submitted as Open Access to the
Journal of the Audio Engineering Society, pp. 1–9, 2016.

To accommodate the smaller format of the printed thesis publication, the layout has been revised. The text is identical to the submitted/published article.

Abstract

In this study the characteristics of compact loudspeakers in a stereo setup were investigated. Perceptual evaluations of eleven loudspeakers were conducted on the basis of six selected sensory descriptors, chosen by experienced listeners during consensus meetings. Based on an analysis of the perceptual evaluation, four of the descriptors were found suited for modelling, with the purpose of developing metrics for prediction of Bass depth, Punch, Brilliance, and Dark-Bright. Bass depth and Punch were modelled as one due to high correlation between them. The experimental setup included loudspeaker spinners, enabling fast positioning of loudspeakers. The prediction models were based on binaural recordings, processed using a loudness model, and developed on the basis of previous work on headphone modelling [1, 2]. They were trained on a subset of the data (66 %) and validated on the rest. The resulting metrics had high correlations with the perceptual ratings of the validation dataset ($r = 0.85$-0.96).

1 Introduction

Loudspeaker specifications have traditionally described the physical properties and characteristics of loudspeakers: Frequency response, dimensions and volume of the cabinet, diameter of drivers, impedance, total harmonic distortion, sensitivity etc. Few of these directly describe the sound reproduction, and none directly describe perception of the reproduction, i.e. takes into account that the human auditory system is highly non-linear in terms of spectral-, temporal- and sound level processing (see e.g. [3]). This disconnect between specifications and perception have made it challenging for acousticians and engineers (and consumers) to predict how a loudspeakers will sound on the basis of these specifications.

Perceptual audio evaluations have long been a reliable method of characterising the reproduction of loudspeakers, headphones, codecs, etc. The requirements for making reliable listening tests are however many, both in terms of facilities, equipment, handling of listeners etc. (see e.g. [4]). Additionally, numerous potential biases [5, 6] must be avoided in the listening test design, making the conduction of listening tests a task for experts only. One way of making perceptual characterisation more accessible (and readily available), have been to develop metrics for predicting perception from various (more easily obtainable) physical measurements of the sound reproduction. The efforts can be divided in two categories: 1) hedonic predictions of e.g. Basic Audio Quality [7], Mean Opinion Score [8, 9], Preference [10] or spatial quality [11], and 2) Predictions of reproduction characteristics, such as Punch [12], Width (sound image) and Bass tightness [13], stereo image width [14], Discolouration, Treble stressing, General bass emphasis, Low bass

emphasis, Brightness, Bass clearness, and Feeling of space [15], and Brightness [16, 17].

While earlier studies focused on making predictions on the basis of the aforementioned specifications (e.g. frequency responses in [18, 19]), more recent modelling efforts have relied more on measurements closer related to the human hearing, e.g. by using binaural recordings as a representation of the physical domain (see e.g. [13, 14]) and by processing the modelling input using auditory models (see e.g. [12–15]).

In the present study the sensory descriptors describing the dominating perceptual differences between compact loudspeakers in a stereo setup were found by consensus meetings with experienced listeners, and predictive models[1], are designed on the basis of listening tests on loudspeakers in a listening room and analysis of binaural recordings made in the listening position. These tests included evaluations of five sensory descriptors[2], representing identified differences on eleven stereo sets of loudspeakers. The loudspeakers were placed in two positions: eight on loudspeaker spinners and three in corner positions. The loudspeaker spinners allowed evaluations of loudspeakers in identical positions with a minimum of switching time, i.e. strain on the limited auditory memory of humans (see review in [21]).

The present study presents a modelling methodology based on binaural recordings being processed using a loudness model. This methodology have been tested for modelling of headphones (sound reproduction with room influences) in a previous study [2]. The present study thereby tests both the suitability of using the proposed methodology for modelling of loudspeakers in a stereo setup, and tests the modelling strategy on a different set of sensory descriptors than previously investigated.

2 Loudspeakers in a stereo-steup

The listening test comprised two sessions; Each with evaluation of seven stereo sets of loudspeakers, of which three sets were in both sessions. The test consisted of reproductions of two musical excerpts evaluated on six sensory descriptors and rated twice by each listener. One session thereby consisted of 168 ratings and had a duration of no more than two hours including breaks, which listeners were encouraged to take whenever needed. The test software automatically regularly reminded listeners to take these breaks. One 'screen' in the test software consisted of evaluation of each of the seven sets of loudspeakers for one sensory descriptors with one musical excerpt,

[1] In this paper the term "metric" is used to describe the end result of the modelling efforts, while "prediction model" is used to refer to the development stages of a "metric".

[2] A sensory descriptor is defined here as a word or phrase that describes, identifies, or labels a perceptual characteristic of a system, e.g. a loudspeaker reproduction. This definition is adapted from [20].

e.g. Bass depth. A 'screen' had seven horizontal rating scales representing each loudspeaker set in a randomised order. The experimental design within one session was a block design with each block consisting of one repetition. Within a block the musical excerpts and sensory descriptors were presented in a randomised order as well. Listeners started both sessions with a familiarisation part that included presentation of all stimuli. In this part, they were allowed to make small adjustments to the overall sound level and instructed to keep that level for the main test.

In the following subsections the details of the setup, the loudspeakers, the stimuli and the listeners are presented.

2.1 Stereo-setup

The listening test was conducted in a listening room compliant with the ITU-R BS.1116-3 [22] recommendation. The loudspeakers were evaluated in the stereo setup depicted in Fig. D.1 (not to scale). Four sets of loudspeakers (spot 1-4) were secured on loudspeaker spinners (DELTA Low Noise Rapid Speaker Spinners), which could move a requested loudspeaker set into the ideal position of the equilateral triangle in about a second no matter the previous position. The figure shows two situations:

Scenario 1 (left) A set of loudspeakers (1-4) on the loudspeaker spinners are playing after being moved into the ideal positions of the equilateral triangle (playback positions).

Scenario 2 (right) A set of loudspeakers in the corners (C1-C3) are playing and the loudspeakers on the spinners are moved to other positions.

Note that the two spinners were always in mirrored positions of each other (not depicted) with two loudspeakers playing in stereo. The loudspeakers on spinners were individually positioned to point towards the listening position (when in the playback positions) and with their acoustically center, as specified by the manufacturers, at 110 ± 0.5 cm above the floor (approximately the height from the floor to the ear canal entrance of an average seated listener). The centre of each loudspeaker spinner was positioned 0.85 m from the side wall and $1.05\,m$ from the back wall. They allowed four sets of loudspeakers to be correctly positioned in an ideal stereo setup (an equilateral triangle), when evaluated by the listeners. Additionally, they are programmed to rotate the least possible (left or right) when moving loudspeakers into the playback positions to minimize switching time. Three additional sets of loudspeakers (C1-C3) where positioned in the corners of the room. A set of Genelec 8020C (C1 positions) were stacked on top of a set of SLA (C2 positions) with Genelec 8050A positioned besides the two (C3 positions). Their acoustic centres were at a height of 145, 122.5 and 133 cm respectively, i.e. higher than the

loudspeakers on the spinners. The set of loudspeakers in the corners were programmed to have individual virtual positions on the loudspeaker spinners. This had two purposes: 1) it rotated the spinners to a position as shown in Fig. D.1 on the right, where the sound emitted was the least obstructed by the loudspeakers on the spinners, and 2) it gave listeners the impression that all loudspeakers were placed on the spinners (important to reduce system identification, which can lead to listener expectation bias [6])

The listener was seated in a chair positioned in the centre of the width-dimension in the room and view of the loudspeakers were blocked by two layers of thin curtains and an acoustically transparent canvas (damping $<$ 1 dB below 16 kHz at an 30° incident angle) displaying the test interfaces.

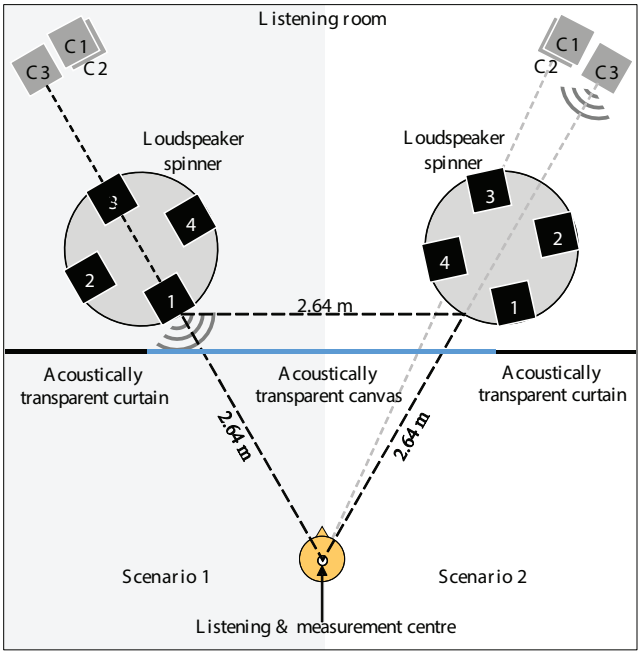


Fig. D.1: Experimental setup. Four sets of loudspeakers were positioned in a standard stereo setup (equilateral triangle) and three other sets in the corners of the room. The loudspeaker spinners move a set of loudspeakers (selected in the test interface) into the playback positions prior to stimuli presentation (Scenario 1). If a set of loudspeakers positioned in the corners (C1-C3) were selected, the spinners instead moved to a position with less influence from the loudspeakers on the spinners (Scenario 2). *Note: The diagram is not to scale.*

2.2 Loudspeakers & Calibration

A perceptual evaluation was made of eleven models of compact loudspeakers listed in Table D.1. Eight were chosen as representative of loudspeakers for in

the consumer segment (price range of 60-554 USD, median of 329 USD) and three were chosen to increase the range of perceptibly differences. Except for one custom-made loudspeaker (SLA), they all had two drivers (tweeter and midrange). The volume of the loudspeaker cabinets were in the range 3-21 l (median 10.4 l), with the exception of the large Genelec 8050A (36 l). The loudspeakers were evaluated in two separate listening sessions to accommodate the space limitations on the loudspeaker spinners. For each session four models were paired to span a wide range of differences between products, i.e. by mixing brands, sizes and price ranges. In the following a *loudspeaker set* refers to two identical loudspeakers used for stereo reproductions.

Eight of the eleven loudspeaker sets were positioned ideally, while three were positioned differently and included in both tests to obtain similar scale usage across the two sessions (as discussed in the previous section). These three are listed in the bottom of Table D.1. The Genelec 8050A loudspeakers were equalised to have an approximately flat frequency response within their specified frequency range, measured in the listening position and with the other four loudspeaker sets on the loudspeaker spinners. The Genelec 8020C were equalised in the same fashion, but had 1/3-octave bands above 10 kHz damped 12 dB to differentiate it from the larger 8050A with regards to both the low- and high frequency extension. Documentation measurements with a pink noise signal showed that the frequency responses of these loudspeakers in the listening position were not identical between sessions (avg. 1/3-octave difference of 2 dB, max. of 12 dB at 12.5 kHz), but the loudspeakers nevertheless received ratings without significant differences between sessions (also indicating that there was no significant session effect in the experimental design). The set of SLA loudspeakers were included in the test to expand the range of perceivable characteristics downwards and were thus not equalised. To slightly reduce the difference to the loudspeakers on the spinners, two strong resonances at 500 Hz and 800 Hz were however dampened 6 dB (using 1/3-octave equalisers).

All loudspeakers were level calibrated to produce 70 ± 0.5 dB(A) in the listening position (measured with a single measurement microphone). The calibration signal had a pink noise spectrum and was band-pass filtered to a frequency range of 80 Hz-14 kHz. After the calibration two of the authors and a colleague checked that no perceptual level differences were noticeable for the chosen stimuli.

2.3 Stimuli & Sensory descriptors

Two musical excerpts were chosen for reproduction over the loudspeakers. A 15 seconds soft pop excerpt (“Bird on a wire” by Jennifer Warnes) and a 24 seconds oriental excerpt (“Moonlight on spring river” by Zhao Cong). Both excerpts were cut to maintain the rhythm during looping. Frequency content

Loudspeaker	Freq. range (±3 dB)	Session
Argon 6340	80 - 20000 *Hz*	2
B&W 685 S2	52 - 22000 *Hz*	2
B&W 686 S2	62 - 22000 *Hz*	2
B&W CM1 S2	50 - 28000 *Hz*	1
DALI Menuet	59 - 25000 *Hz*	1
DALI Zensor 1	53 - 26500 *Hz*	1
DALI Opticon 2	50 - 27000 *Hz*	2
Scandyna MiniPod Mk3	55 - 22000 *Hz*	1
Genelec 8020C	65 - 21000 *Hz*	Both
SenseLab Low Anchor (SLA)	≈80 - 7000 *Hz*	Both
Genelec 8050A	35 - 21000 *Hz*	Both

Table D.1: Selection of compact loudspeakers included in the listening test. Manufacturers' frequency responses are shown. The last three loudspeaker sets were evaluated in both listening sessions and were subject to modifications of their frequency responses.

of the two excerpts are shown in Fig. D.2. The Jennifer Warnes excerpt is dominated by a female vocal and a drum beat, but also includes a variety of other instruments. All sources are clearly separable in the stereo image and the frequency content is smooth in a wide range. The Zhao Cong excerpt is a calm instrumental composition dominated by very deep bass drums and a melody played on pipa (Chinese "lute"). The mix includes many additional instruments as well and has great clarity. The frequency content is broad, but with a lower level in the high bass/low midrange.

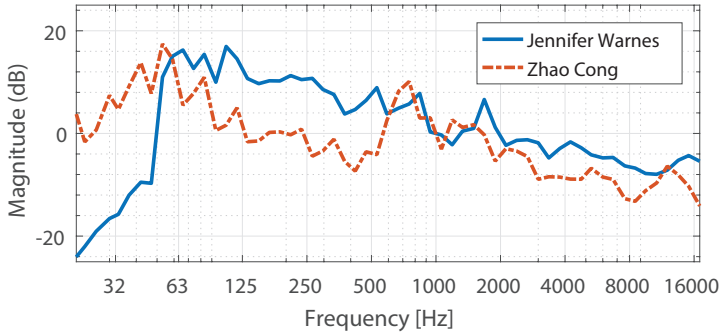


Fig. D.2: Frequency content, L_{EQ}, of the raw musical excerpts depicted in 1/6-octave bands. Normalised to 0 dB at 1 kHz.

Reproduction of the two excerpts were evaluated by listeners on six perceptual characteristics defined by the following sensory descriptors: 1) Punch, 2) Bass depth, 3) Brilliance, 4) Dark-Bright, 5) Natural, and 6) Spatial precision. The descriptors were all from a Sound wheel for audio reproduc-

tion [20] that each have a definition as well as a low- and high verbal scale anchor, e.g. 'a little' and 'a lot' for Brilliance. Note that Punch is defined differently in the Sound wheel compared to that of [12], where it is referred to as something that: *"characterize music or sound sources that convey a sense of dynamic power or weight to the listener"*. Their idea of Punch seemed to have more in common with the descriptors referred to as Bass precision and Attack in the Sound wheel [20]. Punch as defined in the Sound wheel [20] is *"ability to effortlessly handle large volume excursions with compression"*.

2.4 Listeners

Ten listeners participated in the listening test. They were all experienced and trained listeners with normal hearing and ranged in age from 20 to 46 with a median of 30 years. Nine of the ten were trained specially in perceptual loudspeaker evaluation. Their performance was evaluated using a combination of eGauge [23] and Tucker-1 plots. This performance evaluation was described in detail in [2] and includes criteria for removing listeners performing below specified requirements with regards to discrimination and reproducibility.

3 Perceptual modelling & results

3.1 Data basis for perceptual modelling

Out of the six evaluated sensory descriptors, listeners were not able to discriminate between the loudspeakers for Spatial precision and Natural, i.e. none of the loudspeakers on the loudspeaker spinners were rated significantly different from each other on an $\alpha = 0.05$ level. Furthermore, Bass depth and Punch were highly correlated ($r^2 = 0.85$). Consequently, only 'BassPunch' (treating Bass depth and Punch as replicates of the same descriptor), Brilliance, and Dark-Bright were modelled. For each of these, metrics are proposed on the basis of listening test data and corresponding binaural recordings made in the listening position.

The recordings captured the two musical excerpts when reproduced over the loudspeakers. A Brüel & Kjær 4100 head- and torso simulator without ear canals was placed in a chair in the listening position with the microphones centred in a height of 110 ± 0.5 cm above the floor.

The dataset of perceptual ratings and recordings was split up into a training- and validation set. The selection of loudspeaker for each set was chosen separately for each sensory descriptor following this strategy: A first step was to discard perceptual data of the loudspeakers in the corners that had the largest confidence intervals, i.e. session 1 or session 2 data of the same loudspeaker set. This was needed as the loudspeakers were too similar to

treat as separate data points and would have led to overly optimistic evaluation of the prediction models. A second step was to sort the perceptual data per loudspeaker set in ascending order of mean rating and select loudspeakers ranked 2, 5, 8, and 10 for validation ($\approx 36\,\%$) and the rest for training. This selecting scheme ensured that validation ratings were within the range spanned by the training ratings (of which the scope of the prediction models are limited). Additionally, it ensured a wide spread of ratings in both the training and the validation set[3].

3.2 Modelling methodology: BassPunch & Brilliance

Bass depth and Brilliance were defined to describe similar concepts: The bass extension and the treble extension respectively. From a perceptual viewpoint the strongest cue in identifying the bass- or treble extension is loudness at the lowest or highest range of frequencies. Punch is considered related to a temporal response of loudspeakers determined by its time constant (also referred to as onset or rise time). In the Sound wheel definitions [20], Punch is described as related to the reproduction of bass and drums and defined as *"ability to effortlessly handle large volume excursions without compression."*, i.e. to fully reproduce the (relative) level of the low-frequency content. Since the lowest frequencies requires the most of the loudspeakers the correlation with perceived bass depth, seen in the perceptual dataset, seems reasonable. Consequently, both the combined descriptor BassPunch and Brilliance were modelled using a generic methodology proposed in two previous papers for use with perceptual modelling of headphone differences [1] and characteristics [2].

The methodology is based on specific loudness estimations of binaural recordings, here calculated using the time-varying model by Glasberg and Moore [24]. Briefly described, the loudness model corrects for outer- and middle ear influences, calculates the excitation patterns of the basilar membrane, and estimates the specific instantaneous loudness for each millisecond in a frequency resolution of 0.25 equivalent rectangular bands (ERBs). In a final step short- and long term loudness is estimated (taking temporal masking into account). This processing step was, however, not relevant for this purpose as specific loudness was of interest, i.e. averaging over time instead of frequency.

Prediction models were trained using an optimization routine (also described in [1, 2]) that optimized the variables of an equation on the form described by Eq. C.2, such that *metric* correlates the most with the ratings

[3]Note that division of datasets into training and validation subsets is normally done using random draw, i.e. randomly assigning data to one or the other subset, which minimizes the risk of biased/boosted result. With a small dataset this approach isn't suitable as the random subsets risk only spanning a small fraction of the rating range.

of the sensory descriptors. $Dens_m(f)$ is the temporal mean of the instantaneous specific loudness, while A and B denotes the frequency limits of an AB range. The optimization routine searches for the optimum AB range in steps of 0.25 ERBs for A and B independently, but limited to a minimum AB range of 2 ERBs. This limitation was added to reduce the risk of finding spurious high correlations in narrow AB ranges, unlikely to have significantly affected perception and rating of any of the sensory descriptors.

$$metric = \frac{\text{AB range}}{\text{Full range}} = \frac{\sum_{f=A}^{B} Dens_m(f)}{\sum Dens_m} \tag{D.1}$$

In [1], where the methodology was first described, an additional equation was suggested, which had a limited range in the denominator 'CD' as well, as opposed to the full-range of Eq. C.2. This equation was also tested in the present study for modelling BassPunch and Brilliance, but did not lead to as high correlations as the simpler equation in Eq. C.2, and results are consequently not reported.

Additionally the search ranges (investigated AB ranges) were limited to sensible ranges in relation to the general meanings/definitions of bass and treble, namely 20-500 Hz for the BassPunch prediction model and 6.0-14.7 kHz for the Brilliance (14.7 kHz being the highest centre frequency of the loudness model output).

3.3 Modelling methodology: Dark-Bright

In a previous study [2] we described a metric for prediction of Dark-Bright ratings. This metric was based on finding the spectral centroid of the stimuli. While this had been done previously for a descriptor referred to as 'Brightness' (similar in description to Dark-Bright) the novelty was to base the metric on specific loudness estimates instead of frequency content. The metric thereby constitutes the centre frequency at which the loudness in the low and high frequencies are equal (or in practise have minimum difference). Equation 2 describes the solution to the minimization problem of finding the perceptual centroid[4]. $Dens_m(f)$ is again the temporal mean of the instantaneous specific loudness and f is the frequency. f_{MIN}, f_{CEN}, f_{MAX} are the minimum, centroid, and maximum centre frequencies respectively. f_{CEN} thereby

[4]Note that the frequency resolution is 0.25 ERBs and that the total number of frequency bins, 153, was uneven.

represents the point of equal loudness, i.e. the perceptual spectral centroid.

$$\begin{aligned} & f_{CEN}: \\ & \min_{f_{CEN}\in\mathbb{Z}} \quad \left| \sum_{f=f_{MIN}}^{f_{CEN}} Dens_m(f) - \sum_{f=f_{CEN}+1}^{f_{MAX}} Dens_m(f) \right| \\ & \text{subject to} \\ & \qquad f_{MIN} \leq f_{CEN} \leq f_{MAX} \end{aligned} \tag{D.2}$$

As it was hypothesized that the loudness of the mid-frequencies may not influence the perception of the Dark-Bright balance to the same extent as the loudness in the bass- or treble frequency ranges, an alternative prediction model is proposed here. Loudness in the midrange frequencies - defined here to be the range 400-4000 Hz - was reduced in steps of one percent point from $p = 0\,\%$ to $p = 100\,\%$ of the original loudness level $Dens_m(f)$ to investigate the effect on the correlation level with the ratings of Dark-Bright. The optimum value of p, leading to the highest correlation with the perceptual data, was found on the training data and tested on the validation data. Note that this alternative can be viewed as applying a weighted upside-down rectangular window to the specific loudness spectrum, which is unlikely to occur in the human auditory processing. This is, however, a method of testing whether the hypothesis of a weighting function might be part of listeners auditory processing, when evaluating spectral balance. Eq. 2 can be reused for this alternative approach, simply by replacing $Dens_m$ with $Dens_{mw}$, defined in Eq. D.3. Results for the two proposals are reported in Section 3.

$$\begin{aligned} Dens_{mw}(f) &= Dens_m(f) \cdot w(f) \\ &\text{where} \\ w(f) &= \begin{cases} p & \text{for } 0.4 < f < 4.0\,kHz \\ 1 & \text{otherwise} \end{cases} \end{aligned}$$

3.4 Modelling methodology: Logistic transformation

In an effort to obtain models with meaningful predictions in the entire rating interval, i.e. outside the interval of the currently collected data, all prediction models presented so far were transformed using a logistic (s-curve) fit. This ensures that the prediction models can never be outside the range of the scale in the listening test, i.e. 0-15. A logistic transformation ensures a saturation of the prediction value at the lowest and highest end of the scale. The transformation was done in five steps. First, the output of the prediction models was standardized. This was done separately for each of the two musical excerpts to remove excerpt-specific shift and scaling effects. Secondly, a linear fit was used to convert the output to the original rating scale (0-15). This was needed, because the third step required strictly positive values. Thirdly,

a logit transformation was applied (Eq. D.3). This step transforms the data, such that the output and the perceptual data have an approximately linear relationship. Fourthly, a linear fit was found for the logit transformed data. Finally, the linear fit coefficients, $c1$ and $c2$ from step four, were transformed back to the original scale using the logistic transformation (the inverse of the logit transform) in Eq. D.4. The two linear fits (step 2 and 4) were made with perceptual ratings in the training subset and the coefficients were used for both the training and the validation subsets. The results prior to the final step of logistic fitting are depicted in Fig. D.3 to Fig. D.5 with the logistic transformation curve (step 5) plotted.

$$x_2 = \log\left(\frac{x}{15-x}\right) \tag{D.3}$$

$$x_3 = \frac{15}{1+e^{-c1\cdot x_2 - c2}} \tag{D.4}$$

4 Modelling results

The modelling led to metrics for prediction of the sensory descriptors BassPunch, Brilliance, and Dark-Bright respectively. The performance of these are presented in Table D.2 with parameters specified in the Details column. The numbers presented are the Pearson correlation coefficients, r, of the logistic transformed metrics and the AB-ranges are described by their 0.25 ERB centre frequencies (f_c). Scatterplots are depicted in Fig. D.3 to Fig. D.5. Note that for clarity only confidence intervals (CIs) for the validation data set are depicted. The CIs for the training data are similar in size (as both are based on ratings by the same number of listeners).

The big difference between the training and validation coefficients for BassPunch are caused by two sets of loudspeakers being outliers. The fit between perceptual ratings and prediction model prior to logistic transformation is depicted in Fig. D.3. The two outliers (lowest filled blue symbols) are SLA and Genelec 8020C from the corner positions evaluated on the same musical excerpt (Jennifer Warnes). The low correlation coefficient for Dark-Bright (without window) was also caused by two of the loudspeakers in the corners being outliers. The proposed alternative Dark-Bright metric with a weighted upside-down rectangular window, Dark-Bright (R), led to better correlations than the original Dark-Bright metric for both the training- and validation sets.

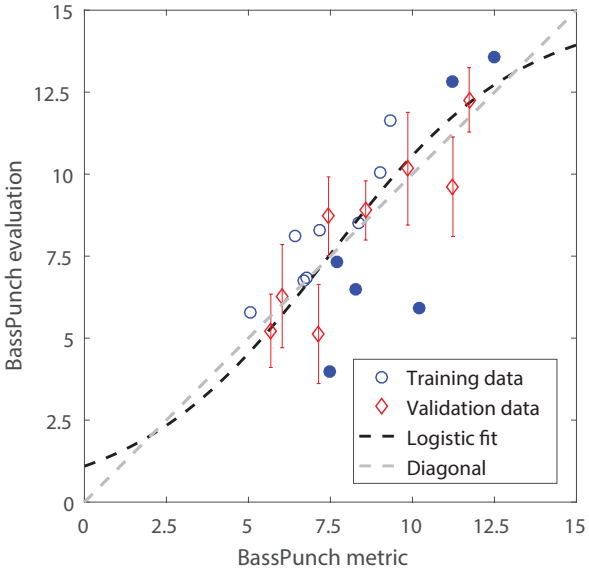


Fig. D.3: Perceptual ratings of BassPunch vs. the proposed prediction model (prior to logistic transformation). Each set of loudspeakers is represented by two data points - one for each musical excerpt. The filled symbols represent loudspeakers from corner positions. The vertical bars represent the 95 %-confidence intervals.

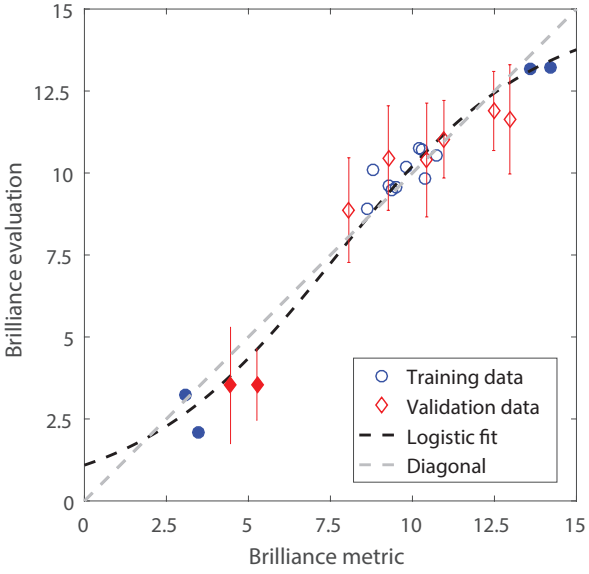


Fig. D.4: Same as Fig. D.3, but for Brilliance.

Metric	Train.	Val.	RMSE	Details
BassPunch	0.70	0.90	1.06	AB: 20-72 Hz
Brilliance	0.99	0.96	1.00	AB: 8.3-10 kHz
Dark-Bright	0.61	0.17	2.19	
Dark-Bright (R)	0.88	0.85	1.08	$p = 7\%$

Table D.2: Performance of metrics describing sensory descriptor ratings. Training (Train.) and Validation (Val.) values are Pearson correlation coefficients, r. RMSE is the root-mean-square error. For the Dark-Bright metrics, (R) denotes the alternative version with a weighted upside-down rectangular window.

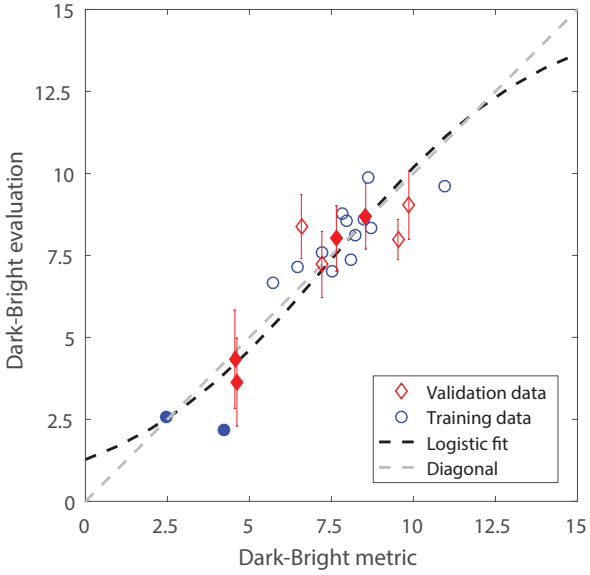


Fig. D.5: Same as Fig. D.3, but for Dark-Bright (R).

5 Discussion

In general, the three proposed metrics, BassPunch, Brilliance and Dark-Bright (R), performed well having correlations coefficients $r \geq 0.85$ for the validation data sets and root-mean-square errors of $RMSE \approx 1$. Furthermore, the AB-range for BassPunch seems intuitively reasonable, while for the Brilliance metric the $10\,kHz$ upper limit seems low and cannot be explained by lack of frequency content in the musical excerpts. The BassPunch metric, however, had a lower correlation coefficient for the training set ($r = 0.70$) than the validation set ($r = 0.90$), and consequently $r = 0.70$ may be the most realistic estimate of its prediction performance level. The outliers in the validation training set are likely to be a consequence of the lack of spectral content at

the lowest frequencies (< 50 Hz) of the Jennifer Warnes excerpt (see Fig. D.2), which constitutes a significant part of the metric's AB range (20-70 Hz).

Due to the understanding of Punch as a characteristic linked to temporal properties of the reproduction, several alternative modelling schemes were tried in an effort to include this aspect in either a separate Punch metric or in a combined BassPunch metric, but none led to consistent predictions. One tested aspect was for instant a modification of Eq. C.2, where the temporal mean of the instantaneous specific loudness, $Dens_m$, was replaced by either the temporal *maximum*, $Dens_{max}$ or the mean of an upper percentile, e.g. the 90th, of the instantaneous specific loudness, $Dens_{mp}$.

The higher performance of the Dark-Bright (R) metric in comparison to the simpler Dark-Bright metric proposed in [2], suggests that listeners put less emphasis on the midrange frequencies, when evaluating the spectral balance of sound reproduction. It is, however, important to reiterate that the proposed weighting function is unlikely to correspond to that of a listener, as it has two strong discontinuity points at the start- and end frequencies of the function. A smoother function is expected to better represent this step in the auditory processing. Additional research is also required to establish the best-fit frequency limits, as the current range, 400-4000 Hz, were chosen only as an initial estimate. Furthermore, it is of interest to establish whether these limits are similar for all listeners or whether clusters exists. Investigation of these improvements are planned to be the subject of a future study.

In terms of performance of the proposed metrics, it was of further interest whether all confidence intervals of the validation data points overlap the curve of the logistic fit in Fig. D.3 to Fig. D.5, in which case the best possible fit is reached within the uncertainty of the data and more data would be needed to verify further improvements. The Brilliance metric reached this prediction performance level, while small improvements are still possible within the statistical uncertainties of the current data set for both the BassPunch and the Dark-Bright metrics.

In comparison with previous modelling effort in the literature, one important difference in this study, is the definition of the listeners "internal reference. The traditional view is for instance seen in [15], where Klippel proposed seven metrics for describing loudspeaker performance. The basis of his metrics was a calculation of "discolouration", which were defined as stated in Eq. D.5, where $N'_{test}(z)$ and $N'_{ref}(z)$ are the specific loudness of the test- and reference stimuli respectively.

$$N'(z) = N'_{test}(z) - N'_{ref}(z) \tag{D.5}$$

Eq. D.5 implies that the listeners know the recorded reference and are able to use this as an "internal reference" for assessment of loudspeakers by the deviations from this reference. The weakness here is that the listener does not

know the recorded reference, as it cannot be presented to the listeners without being affected by the reproduction system. This approach is also used in prediction models involving codecs, e.g. P.OLQA [9] and QESTRAL [11], but here the discolouration of the reproduction system is included in both the reference and the compressed systems under test. In the present study, the perceptual sound reproduction characteristics were defined as: *"The perceived changes to the envisioned original sound"*. So, we assumed that the listener creates an internal reference of the original sound on the basis of what is heard and assess the loudspeaker characteristics as the deviation from this reference. The weakness here is that the internal reference is dependent of both the characteristics of the musical excerpts and the loudspeakers under evaluation. We try to overcome this weakness by letting the assessors listen to all systems with different musical excerpts (familiarization, see Section 2) before the listening test, such that the internal reference should be an average over excerpts and thereby be a tool for assessing the loudspeakers with limited influence of the specific excerpts. In the processing of data, the internal reference in the present study was approximated by averaging over stimuli available in the training set and used for standardization as described in Section 3.4 (step 1).

Besides the different definitions of listener reference, the study by Klippel [15] showed many similarities supporting the findings of the current study. His metrics were based on seven sensory descriptors identified by comparisons of loudspeakers using a combination of ratio- and multidimensional scaling methods. They were analysed using factor analysis and thereby comprise a list of dominating perceptual differences between the sound reproduction of loudspeakers. Four of these are similar to the set used in the present study, i.e. 1) Treble Stressing $\approx$ Brilliance, 2) Low bass emphasis $\approx$ Bass depth/Punch, 3) Brightness $\approx$ Dark-Bright and 4) Feeling of space $\approx$ Spatial precision. Note, however, that Klippel's Treble Stressing was linked to the perception of sharpness or shrillness, where Brilliance is defined as treble extension. Klippel's proposed metric Low bass emphasis describes the ratio between the discolouration below $f_c = 60$ Hz and all critical bands above, with discolouration defined as spectral deviation from the original stimulus (discussed above). This is comparable to the AB-range found for BassPunch of 20-72 Hz (see Table D.2).

Klippel's proposed Brightness (Dark-bright) metric is shown in Eq. D.6, where S is Treble stressing metric and B is General bass emphasis.

$$H = 0.7S - 0.3B \tag{D.6}$$

B is calculated from the same equation as Low bass emphasis, but with a pivot point at $f_c = 150$ Hz. S is based on discolouration as well, but multiplied by a weighting function increasing at the higher frequencies. Consequently, Klippels Brightness metric puts higher emphasis on bass and treble

than on midrange frequencies as well, but additionally puts higher weight on treble than bass, which may be a consequence of a low pivot point at 150 Hz, which does not encompass the full bass frequency range.

6 Summary

In this study, three metrics were developed for prediction of the perceived characteristics of loudspeakers' sound reproduction in a stereo setup with regards to BassPunch, Brilliance and Dark-Bright. The metrics were developed with the intention of finding specifications of loudspeakers' sound reproduction with perceptual relevance. They were based on binaural recordings made in the same setup, as was used for perceptual evaluations of eleven stereo sets of loudspeakers. The recordings, made using a head- and torso simulator, were processed using a loudness model and led to metrics describing spectral characteristics of the reproduction. Two, were based on the relative specific loudness of a limiting frequency range (AB) and one was based on a weighted specific loudness centroid. The prediction models were trained on a training subset with seven sets of loudspeakers and validated on four others. The range of correlation coefficients were $r = 0.85$-0.96 (details in Table D.2, page 127). All metrics thus showed potential for prediction of a comparable loudspeaker segment and with a root-mean-square-error of $RMSE \approx 1$ on a 0-15 rating scale for the validation set. This RMSE level was largely comparable to the statistical 95 % confidence intervals of the perceptual evaluations.

Acknowledgement

This work was funded by DELTA and the Danish Agency for Science, Technology and Innovation (Case number: 1355-00061). The authors wish to thank Hi-Fi Klubben for making eight of the compact loudspeakers available for this study as well as Tore Stegenborg-Andersen for assistance in conducting the listening test.

References

[1] C. P. Volk, M. Lavandier, S. Bech, and F. Christensen, "Identifying the dominating perceptual differences in headphone reproduction," *Submitted, J. Acoust. Soc. Am.*, Feb. 2016.

[2] C. P. Volk, T. H. Pedersen, S. Bech, and F. Christensen, "Modelling perceptual characteristics of prototype headphones," in *Proc. of the 2016 AES International Conference on Headphone Technology*, (Aalborg, Denmark), pp. 1–9, Audio Engineering Society, Aug. 2016.

[3] C. J. Plack, *The sense of hearing*. Mahwah, N.J: Lawrence Erlbaum Associates, 2005.

[4] S. Bech and N. Zacharov, *Perceptual audio evaluation: theory, method and application*. Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2006.

[5] S. Zielinski, "On Some Biases Encountered in Modern Audio Quality Listening Tests (Part 2): Selected Graphical Examples and Discussion," *J. Audio Eng. Soc*, vol. 64, pp. 55–74, Feb. 2016.

[6] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review," *J. Audio Eng. Soc*, vol. 56, no. 6, pp. 427–451, 2008.

[7] ITU-R, "Method for objective measurements of perceived audio quality," Recommendation ITU-R BS.1387-1, International Telecommunication Union Radiocommunication Assembly (ITU-R), United States, 2001.

[8] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Recommendation ITU-T PP.862, ITU Telecommunication Standardization Sector (ITU-T), United States, 1999.

[9] ITU-T, "Perceptual objective listening quality assessment," Recommendation ITU-T P.863, ITU Telecommunication Standardization Sector (ITU-T), United States, Jan. 2011.

[10] S. E. Olive, "A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model," in *Proc. of the Audio Engineering Convention 117*, (San Francisco, CA, USA), pp. 1–21, Audio Engineering Society, 2004. Convention paper 6190.

[11] M. Dewhirst, R. Conetta, F. Rumsey, P. Jackson, S. Zielinski, S. George, S. Bech, and D. Meares, "QESTRAL (Part 4): Test Signals, Combining Metrics, and the Prediction of Overall Spatial Quality," in *Proc. of the Audio Engineering Society Convention 125*, (San Francisco, CA, USA), pp. 1–8, Audio Engineering Society, 2008. Convention Paper 7598.

[12] S. Fenton and H. Lee, "Towards a perceptual model of "Punch" in musical signals," in *Proc. of the Audio Engineering Society Convention 139*, (New York, NY, USA), pp. 1–10, Audio Engineering Society, Oct. 2015. Convention paper 9381.

[13] S. Kim and W. L. Martens, "Deriving Physical Predictors for Auditory Attribute Ratings Made in Response to Multichannel Music Reproductions," in *Proc. of the Audio Engineering Society Convention 123*, (New York, NY, USA), pp. 1–10, Audio Engineering Society, Oct. 2007. Convention paper 7195.

[14] M. Takanen and G. Lorho, "A Binaural Auditory Model for the Evaluation of Reproduced Stereophonic Sound," in *Proc. of the Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*, (Helsinki, Finland), pp. 1–10, Audio Engineering Society, Mar. 2012.

[15] W. Klippel, "Multidimensional Relationship between Subjective Listening Impression and Objective Loudspeaker Parameters," *Acta Acust. united Ac.*, vol. 70, no. 1, pp. 45–54, 1990.

[16] I. B. Labuschagne and J. J. Hanekom, "Preparation of stimuli for timbre perception studies," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2256–2267, 2013.

[17] E. Schubert and J. Wolfe, "Does Timbral Brightness Scale with Frequency and Spectral Centroid?," *Acta Acust. united Ac.*, vol. 92, pp. 820–825, 2006.

[18] A. Gabrielsson, "Perceived sound quality of reproductions with different frequency responses and sound levels," *J. Acoust. Soc. Am.*, vol. 88, no. 3, p. 1359, 1990.

[19] F. E. Toole, "Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2," *J. Audio Eng. Soc*, vol. 34, no. 5, pp. 323–348, 1986.

[20] T. H. Pedersen and N. Zacharov, "The Development of a Sound Wheel for Reproduced Sound," in *Audio Engineering Society Convention 138*, (Warsaw, Poland), pp. 1–13, Audio Engineering Society, May 2015. Convention Paper 9310.

[21] N. Cowan, "On short and long auditory stores," *Psychol. Bull.*, vol. 96, no. 2, pp. 341–370, 1984.

[22] ITU-R, "Recommendation BS 1116-3, Methods for the Subjective Assessment of Small Impairments in audio Systems Including Multichannel Sound Systems.," Recommendation ITU-R BS 1116-3, International Telecommunication Union Radiocommunication Assembly (ITU-R), United States, Feb. 2015.

[23] G. Lorho, G. Le Ray, and N. Zacharov, "eGauge—A Measure of Assessor Expertise in Audio Quality Evaluations," in *Proc. of the Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, (Piteå, Sweden), pp. 1–10, Audio Engineering Society, June 2010. Paper No. 7-2.

[24] B. R. Glasberg and B. C. J. Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *J. Audio Eng. Soc*, vol. 50, no. 5, pp. 331–342, 2002.

Part III

Appendix

Characterisation of acoustical environments: Physics and perception

Abstract

Fifteen acoustical environments were characterised with respect to a number of electro-acoustical measurements in the physical domain and a number of sensory descriptors in the perceptual domain. The physical measurements were made in the same position as the listeners were situated during the perceptual evaluation to allow for a direct comparison of the two domains. The data quality analysis of the perceptual data revealed problems in the experimental design and setup, which are described and discussed in detail in this appendix.

Practical application
The methods of data quality analysis and the problems discovered may help others avoid a number of pitfalls with regards to perceptual evaluations of audio reproduction equipment.

Project background
The purpose of this investigation was two-fold: Firstly, to collect perceptual data for modelling the relation between the physical acoustical environment and the human perception (PhD project activity). Secondly, to evaluate the suitability of the included sensory descriptors (DELTA SenseLab activity).

1 Introduction

Two listening tests were conducted in an effort to investigate the general question of how differences in an acoustic environments driven by reproduced audio systems are perceived and evaluated by listeners. This was investigated through listening tests and measurements of the electro-acoustical properties of the acoustical environment in the listening position. Examples of measurements include physical frequency response measurements as well as perceived Bass strength and Naturalness evaluations.

The perceptual evaluations took place in two iterations: Eight systems were evaluated in a first test (Test 1) and ten systems were evaluated in a second test (Test 2). Each system included sound reproduction of a loudspeaker, but the setup added significant influences to the sound reaching the listening position. Consequently, a strict distinction is made in this appendix between a loudspeaker and a *system* (defined later on). Three loudspeakers from Test 1 were included in Test 2 to allow for a potential data aggregation of the two datasets, leading to a total of $8 + (10 - 3) = 15$ systems. The three recurring loudspeakers were assumed to provide the same perceived sound at the listening position in both setups. This assumption is based on the loudspeakers being positioned in the same positions and being influence very similarly by the two setups. On the basis of the results from Test 1, the loudspeakers in Test 2 were selected to represent a wider perceptual intensity range for each of the included sensory descriptors, as the systems in Test 1 were found to have too low variability, i.e. spanning a limited part of the rating scales.

In these two tests a number of loudspeakers were used to create variation in the acoustical environment surrounding the listeners. This was *not* an eval-

Fig. I: The "baffle" setup of loudspeakers for Test 2. The white-framed loudspeaker (lower left) was kept for symmetry, but only the right was included in the test. The large Genelec (leftmost) was placed on top of the white left loudspeaker in Test 1. The position of the two other anchor loudspeakers were fixed between Test 1 and Test 2.

uation of loudspeakers, but of loudspeaker-driven acoustical environments. The loudspeakers were positioned up against each other in a "baffle" in an effort to minimize spatial cues that the listeners could have used to iden-

tify each individual loudspeaker. The "baffle" setup from Test 2 is depicted in Fig. I. The loudspeakers were placed on a table to position the drivers in a height, which were similar to that of a listeners' ears above the floor, to preserve the high frequency components and the frequency reproduction balance in the direct sound.

As a result of this setup, each of the audio reproduction systems in these tests were, from the listeners perspective, a combinations of a loudspeaker output, influence from presence of surrounding loudspeakers, influence from position of the loudspeaker, room influence, and all interactions between these elements.

Electro-acoustic measurements of all 15 systems were made at the listening position. Making the measurements of these systems in the listening position where the listeners were positioned during the listening test, allowed a one-to-one investigation of how perceived differences from one system to another were related to differences in the electro-acoustical measurement data.

2 Measurement methodology

2.1 System Description

The 15 loudspeakers used in the two tests are presented in Table I. The loudspeakers were selected by colleagues through several informal listening sessions, focusing on getting a selection of loudspeakers that would lead to audible differences between systems with respect to the sensory descriptors included in the two listening tests.

2.2 Listening room setup

A digram of the listening room setup is depicted in Fig. II. The listening room is in compliance with the ITU BS.1116-1 Recommendation and EBU 3276 standard regarding listening rooms, with a low reverberation time, a floor area of $4.69\,m \times 7.84\,m = 36.80\,m^2$ and an volume of $96.43\,m^3$. The "baffle" of loudspeakers were placed behind an acoustical transparent screen, severing the dual purpose of displaying the test user interface and hiding the loudspeakers, such that visual bias is avoided.
The listening position was situated with equal distance to the left and right wall, but further from the front wall (4.74 m) than the back wall (3.1 m). The longer distance from the "baffle" to the listening position reduced the difference in azimuth between sources, making identification by localization more difficult.

Manufacturer	Model	Type	Crossover	Outer vol.	Description
Test 1 systems					
DALI	Epicon 2	2-way	3.1 kHz	30.3 L	High-end
DALI	Epicon 2				16Ω in series
DALI	Mentor 2	3-way	3.4/12 kHz		Hybrid tweeter
DALI	Menuet	2-way	3.0 kHz	8.6 L	Ultra compact
DALI	Menuet				16Ω in series
Test 2 systems					
JBL	4208	2-way		35.3 L	
Spendor	LS3/5A	2-way		32.1 L	BBC design
Technics	SB-F820	2-way		11.7 L	
Sony	SS CSE1	2-way		11.2 L	
Aiwa	SX-MS7	2-way		3.6 L	
Celestion	CXi 521	2-way		13.3 L	
Tannoy	CPA5	2-way		4.4 L	Coaxial (5",1")
Anchor systems					
Homebuild		2-way		32.1 L	Undamped
Genelec	8020C	2-way	3.0 kHz	4.8 L	Active
Genelec	8050A	2-way	1.8 kHz	35.9 L	Active

Table I: Description of the loudspeakers used in the two listening tests. The anchor systems were included and evaluated in both Test 1 and Test 2.

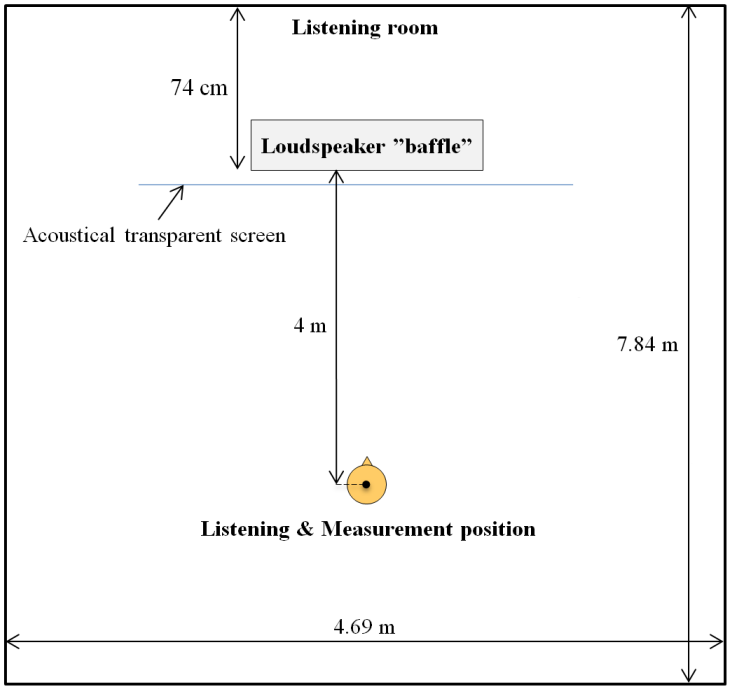


Fig. II: Sketch of the listening room setup. The ceiling height is 2.62 meters. The listeners for the listening test and the equipment for the electro-acoustical measurements where positioned in the same "sweet spot". Note: The sketch is not to scale.

2.3 Setup and equipment

The tests were carried out using a multiple comparison method. The software utilized to run a listening test with this methodology was custom-made and programmed in Labview. The software uses ASIO drivers allowing complete audio control without interference from the operating system (Windows 7). The hardware used in the stimuli reproduction is presented in Table II.

Equipment	**Type**
Computer	Lenovo ThinkCentre Tower
Sound card	RME Fireface 800, 16 ch. (ASIO interface)
D/A ADAT	RME M-32 DA
Power amplifiers	NAD C326BEE, 2x80W (x4)

Table II: Equipment used for stimuli reproduction.

2.4 Calibration and equalisation

The systems were calibrated to a level of $70\,dB(A) \pm 1\,dB$ in the listening position using a band-limited pink noise test signal as described in [1]. This level was chosen as a realistic listening level, which the listener would feel comfortable listening to for two hours. During an informal listening session by three experienced listeners, this level was assessed to be within the linear performance area of the loudspeakers.

2.5 Listening tests: General description

Sensory descriptor group	Stimulus 1	Stimulus 2	Stimulus 3
Timbre: Bass (4)	•	•	
- Midrange (1)	•	•	
- Treble (1)	•	•	
- Spectral balance (2)	•	•	
Dynamics (4)	•	•	
Transparency (4)	•	•	•
Preference	•	•	•

Table III: Listening tests' design. The numbers in parentheses in the first column describes the number of sensory descriptors evaluated within a group of related descriptors from the Sound wheel [2]. A bullet signifies that the row-column combination was included in the tests.

The two listening tests were based on the same physical setup, the same sensory descriptors, design of experiment, procedure, and instructions. The listening tests' design are presented in Table III.

The three stimuli were musical excepts of durations of $18 - 23$ s from: 1) Shirley Caesar - Stand Still (SC) [3], 2) Paula Cole - Tiger (PC) [4] and 3) Eliane Elias - Chega De Saudade (EE) [5]. The Shirley Caesar excerpt has a female vocal, multiple drums (transients) and a dominating trumpet; The excerpt has an almost flat spectrum in the range from $40\,Hz$ to $8\,KHz$, when measured over the whole excerpt. The Paula Cole excerpt has a female vocal, drums, and a deep dominating five-string bass (lowest note: B0 at $30.868\,Hz$). The Eliane Elias excerpt has a female vocal, and an acoustic guitar, and was specifically chosen for evaluation of transparency and naturalness.

Each combination in the test design shown in Table III were repeated twice during the listening test. The design of experiments (DoE) was a 3-part block design consisting of 1) Preference, 2) Timbre & Dynamics, and lastly 3) Transparency. Systems and musical excerpts were randomised within each block. The six blocks (3×2 repetitions) were structured in three sessions designed to have a duration of no more than 2 hours including instructions and breaks. Listeners did, however, had more time available if needed.

The 16 sensory descriptors were selected on the basis of results from a prior training test. They were considered an experimental set of sensory descriptors, still under investigation with regards to suitability. This is also the reason why more descriptors were included than is commonly done. The sensory descriptors and their English definitions are described in the Sound wheel [6, pp. 21-23]. Their meaning and relevance will not be discussed further in this appendix, it is however important to note, that one sensory descriptor was evaluated different from the remaining 15. For the sensory descriptor Råstyrke (Powerful) from the Dynamics group), listeners were instructed to increase the volume of the reproduction and base their evaluation on the level at which the reproduction became distorted as well as the degree and nature of the distortion.

The evaluation methodology was a multiple comparison scheme, with each screen on the graphical user interface (GUI) displaying all systems for one sensory descriptor. An example is presented in Fig. III. Due to the special instructions for evaluation of the sensory descriptor Råstyrke (Powerful), the interface contains a master volume control and listeners were instructed to gradually increase the volume from the initial $-20\,dB\,FS$ until the distortion became obvious or until a maximum volume of $-10\,dB\,FS$ was reached. The maximum was found needed to avoid damaging the smaller loudspeakers.

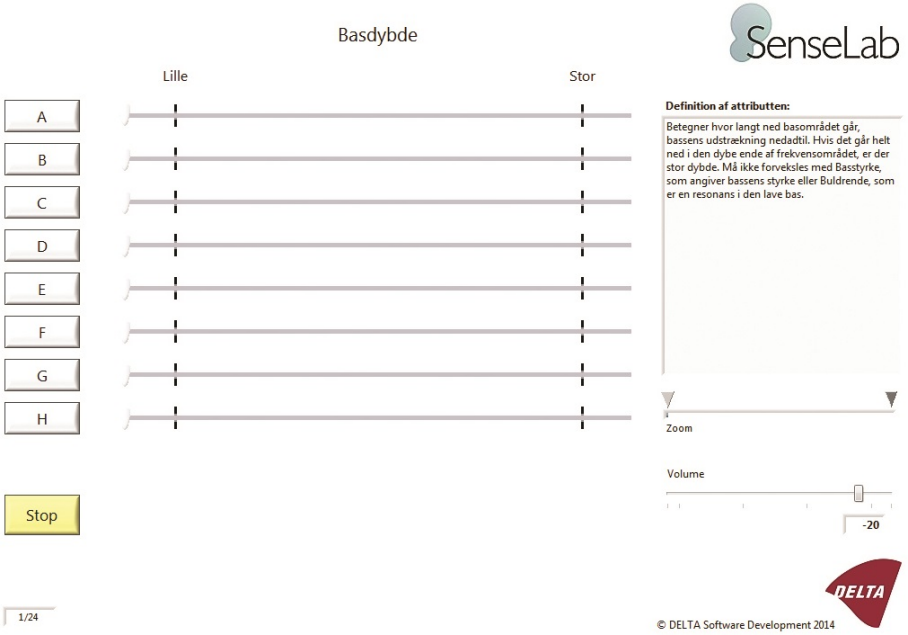


Fig. III: Example of the graphical user interface of the multiple comparision scheme of evaluation used in both perceptual experiments. The distance between graphical components have been reduced for clarity in figure.

2.6 Listeners

Eight listeners from DELTA's panel of experienced listeners were specially selected to be trained specifically in evaluation of loudspeakers. This panel was named specialized expert assessor panel (SEAP) and is referred to as such in this appendix. The eight listeners were part of a long process of selecting a set of sensory descriptors suited for exhaustive/complete evaluation of differences between reproduced audio products and were trained in the use of these sensory descriptors. The sensory descriptor selection process is also described in [2].

3 Listening Test 1

The first listening test included five DALI loudspeakers and three anchors systems (See Table I, page 138). Two of the DALI loudspeakers were modified by putting a 16Ω resistor in series, in an effort to affect the performance of the two systems[5] and increase the variation in stimuli between the test systems. The DALI loudspeakers ranged in price from 3,500 to 16,500 DKK per unit.

3.1 Data Quality analysis 1

The results of the first listening test showed a clear problem with the test. What is presented in this section is only the results needed to understand the problem and not a detailed data analysis. In Fig. IV the average rating of each system for each musical excerpt is plotted. Furthermore, based on a 3-way ANOVA model the least significant difference (LSD) for each sensory descriptor was calculated and is represented by the vertical bars in Fig. IV. If system ratings are within the bar, then the system cannot be statistically proved to be different. An example is the mean system ratings of "Brillians", where only System 5 is outside the LSD bar, implying that only System 5 is statistically different (with $p < 0.001$) from any other system. The position of the bars have been manually placed to cover the largest number of systems, thus visualizing the large similarity in ratings of systems. Notice also that the LSD is shown without any correction, e.g. Bonferroni or Holm, and may thereby underestimate the length of the bars; The consequence begin that less systems would be found significantly different. From Fig. IV it was clear that the evaluated systems were very similar, i.e. a majority of systems were rated within the length of the bar, for all but the three bass sensory descriptors. Furthermore System 5 (Homebuild) was perceived or identified

[5] To give perspective this corresponds to $l = \frac{RA}{\rho} = \frac{16\,\Omega \times 0.823 \cdot 10^{-6}\,m^2}{1.7 \cdot 10{-}8\,\Omega\,m} = 774\,m$ of thin 18 AWG ($0.823\,mm^2$) pure copper cable.

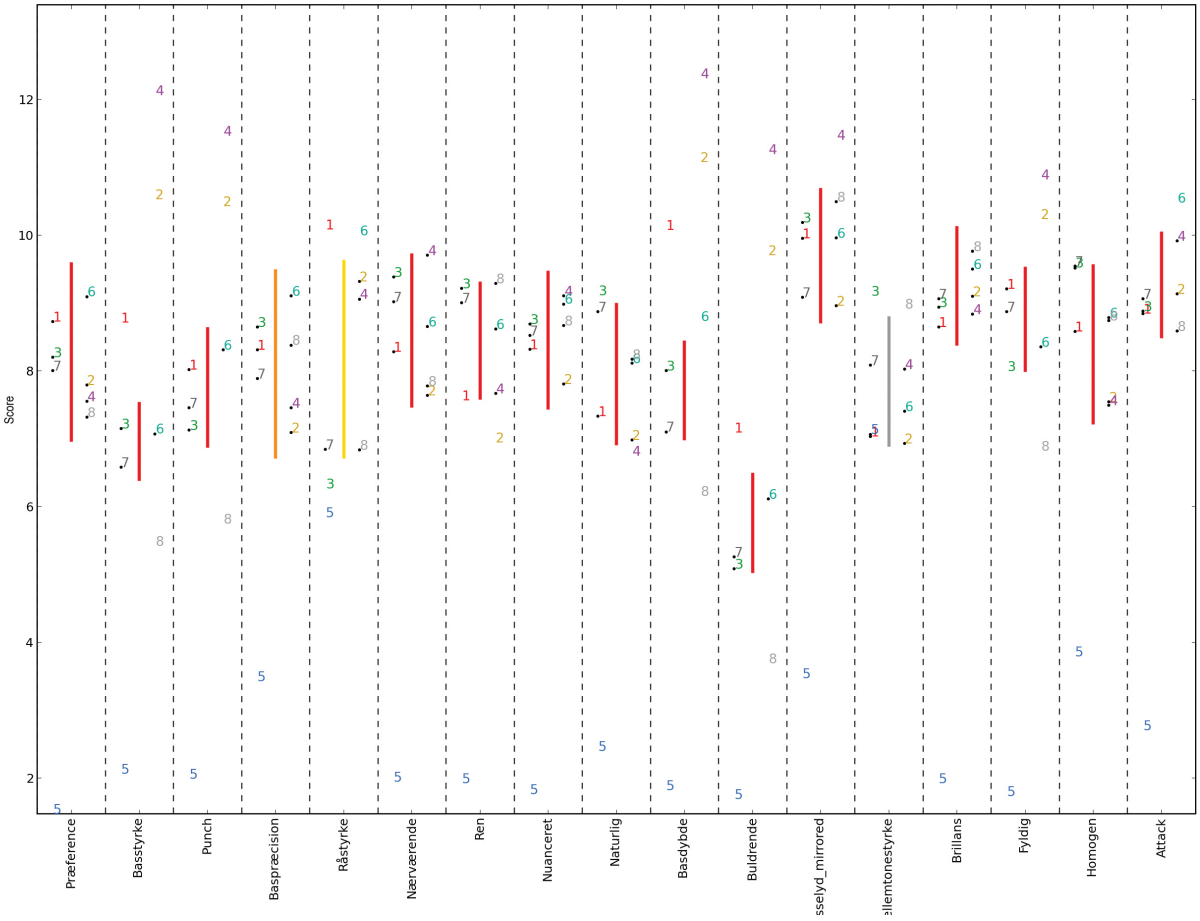


Fig. IV: *Test 1:* Overview of sensory descriptor evalutions. Each number represents a system and the position on the vertical axis represents the mean score on the original rating scale across all listeners and musical excerpts. The vertical bars indicate the least significant difference (LSD) and the color indicate the degreee of significance associated with the length of the bar. Red is $p < 0.001$, orange is $p < 0.01$, yellow is $p > 0.05$ and grey is non-significant. A small black dot next to a number (system) indicates that the bar overlaps with the system's mean rating. The plot was generated using PanelCheck [7].

by the listeners to be a low anchor - always scoring much lower than the other systems (or inverse for Kasselyd).

3.2 Discussion 1

The system with the Homebuild loudspeaker, turned out to be treated by the listeners as a low anchor in the test. This may be due to their prior experience from MUSHRA tests [8], where a low anchor is always included, or it may be due to a real difference in perception between the systems. A combination is also possible. In MUSHRA tests, the low anchor is a low-pass filtered ($3.5\,kHz$) version of the original musical excerpt often given a fixed low rating by listeners (discussed in [9]). In all three scenarios, the low anchor may have affected the scale usage in such a way that the remaining systems were squeezed together in the top-half of the scale, which would decrease the differences in ratings even for systems that the listeners were able to discriminate between. This phenomenon constitutes a type of Contraction bias [10]. This bias is normally associated with a centering of the ratings, in this case it is a contraction of the ratings of the majority of the systems due to the perceived outlier system.

Another observation was, that the majority of systems were rated similarly for all sensory descriptors except Basstyke and Buldrende. This was theorised to stem from having the majority of loudspeakers being from one brand, thus likely to be tuned to the same target frequency response (with the possible exception of the low-frequency performance, where the physical dimensions of the loudspeakers becomes a limitation).

The small span of system ratings and the similar ratings for the majority of sensory descriptors provided a challenge with regards to modelling the link between perception and electro-acoustical measurements. If no differences exists between systems ratings, the relation to the physical system cannot be modelled. Furthermore, if the systems are rated the same for many sensory descriptors, e.g. for perceived bass strength and treble strength, a model predicting bass strength well, would equally well predict treble, which is assumed not to be generally valid, but a trait of the specific data set used for training the model and without causal foundation.

As a result a second test was initiated, which included a larger span of loudspeakers with regards to quality, size, brands, and technology. This test is described in the next section.

4 Listening Test 2

In the second test (Test 2), seven different loudspeakers were included as well as three from the previous test. The three reacurring loudspeakers were included as anchor systems to allow for a merge of the two data sets and were chosen as they spanned a large part of the scale in the previous test (See systems 3,4, and 5 in Fig. IV), i.e. one in the bottom of the scale, one in the middle and one in the high end of the scale. The other seven loudspeakers were generally inferior to the loudspeakers in Test 1, but chosen (by informal listening) as they were believed to span a larger part perceptual space for all of the included sensory descriptors. The loudspeakers in Test 2 are also described in Table I. Besides from the different loudspeakers, the design of experiment was a replicate of the previous test.

4.1 Data quality analysis 2

In Fig. V the LSD plot is presented for the data from Test 2. Comparing with Test 1, the spread of system ratings are now much larger and spanning the majority of the scale. The LSD bars indicates that differences exists between systems with a very high probability ($p < 0.001$), which is a consequence of the the larger spread in ratings and of systems being evenly distributed on the scale.

In Fig. VI the correlation matrix showing the Pearson Product-Moment correlations are depicted. Cells coloured red have absolute correlation values $|r| \geq 0.87$, meaning that at least $r^2 = 0.87^2 = 75\,\%$ of the variance is explained by the other sensory descriptor. This is evidence of a strong dependence among the sensory descriptor ratings.

Four sensory descriptors stood out from the analysis: Buldrende, Baspræcision, Råstyrke and Mellemtonestyrke. Investigating Fig. VI, it is clear that Mellemtonestyrke is rated neutral for all systems, i.e. even through the sensory descriptor is different from the other sensory descriptors it has no power to differentiate the systems, and thus no prospect for modelling. As for Buldrende and Baspræcision, the correlation coefficients are still high and all sensory descriptors except from Mellemtonestyrke (and Råstyrke for Baspræcision) can explain more than 50 % of its variance and all correlations are moreover still significant on a $p < 0.001$ level.

Another view of the dataset is visualised in the PCA plot in Fig. VII. The first dimension explains 54.6 % of the variance in the dataset. From the scree plot (right) it is clear that the knee point is between the first and second dimension. From the Loadings plot (left) it can be observed that the ratings have a high overlap between the sensory descriptors and that only

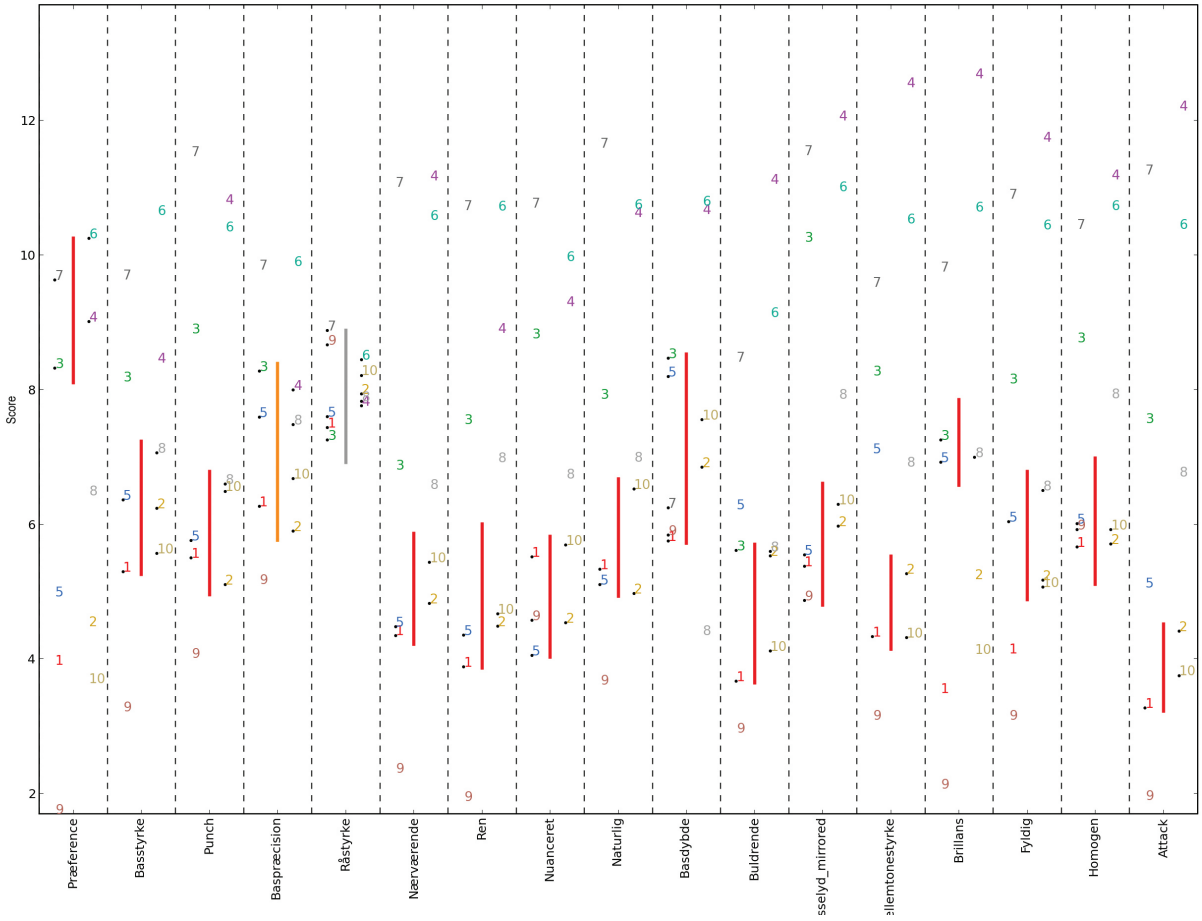


Fig. V: *Test 2:* Overview of sensory descriptor evalutions. Each number represents a system and the position on the vertical axis represents the mean score on the original rating scale across all listeners and musical excerpts. The vertical bars indicate the least significant difference (LSD) and the color indicate the degreee of significance associated with the length of the bar. Red is $p < 0.001$, orange is $p < 0.01$, yellow is $p > 0.05$ and grey is non-significant. A small black dot next to a number (system) indicates that the bar overlaps with the system's mean rating. The plot was generated using PanelCheck [7].

	Basstyrke	Basdybde	Buldrende	Fyldig	Punch	Attack	Nærværende	Nuanceret	Brillans	Homogen	Naturlig	Præference	Ren	Baspræcision	Råstyrke	Mellemtonestyrke	Kasselyd
Basstyrke	1	0,99	0,98	0,97	0,97	0,92	0,93	0,89	0,9	0,89	0,89	0,92	0,8	0,83	0,66	0,04	-0,91
Basdybde	0,99	1	0,97	0,97	0,97	0,93	0,92	0,89	0,91	0,88	0,89	0,93	0,82	0,83	0,69	-0	-0,92
Buldrende	0,98	0,97	1	0,95	0,95	0,89	0,92	0,86	0,86	0,84	0,84	0,86	0,75	0,76	0,69	0,11	-0,86
Fyldig	0,97	0,97	0,95	1	1	0,97	0,98	0,96	0,96	0,91	0,95	0,96	0,9	0,87	0,57	0,15	-0,97
Punch	0,97	0,97	0,95	1	1	0,96	0,98	0,96	0,97	0,92	0,95	0,96	0,9	0,88	0,63	0,15	-0,97
Attack	0,92	0,93	0,89	0,97	0,96	1	0,95	0,95	0,95	0,87	0,93	0,94	0,95	0,84	0,56	0,23	-0,98
Nærværende	0,93	0,92	0,92	0,98	0,98	0,95	1	0,99	0,97	0,92	0,97	0,94	0,93	0,87	0,55	0,23	-0,96
Nuanceret	0,89	0,89	0,86	0,96	0,96	0,95	0,99	1	0,99	0,92	0,97	0,94	0,96	0,9	0,52	0,27	-0,96
Brillans	0,9	0,91	0,86	0,96	0,97	0,95	0,97	0,99	1	0,91	0,96	0,95	0,96	0,91	0,56	0,2	-0,97
Homogen	0,89	0,88	0,84	0,91	0,92	0,87	0,92	0,92	0,91	1	0,97	0,98	0,87	0,95	0,55	0,11	-0,89
Naturlig	0,89	0,89	0,84	0,95	0,95	0,93	0,97	0,97	0,96	0,97	1	0,98	0,94	0,94	0,48	0,22	-0,95
Præference	0,92	0,93	0,86	0,96	0,96	0,94	0,94	0,94	0,95	0,98	0,98	1	0,92	0,94	0,55	0,06	-0,95
Ren	0,8	0,82	0,75	0,9	0,9	0,95	0,93	0,96	0,96	0,87	0,94	0,92	1	0,87	0,44	0,25	-0,96
Baspræcision	0,83	0,83	0,76	0,87	0,88	0,84	0,87	0,9	0,91	0,95	0,94	0,94	0,87	1	0,48	0,18	-0,85
Råstyrke	0,66	0,69	0,69	0,57	0,63	0,56	0,55	0,52	0,56	0,55	0,48	0,55	0,44	0,48	1	-0,09	-0,56
Mellemtonestyrke	0,04	-0	0,11	0,15	0,15	0,23	0,23	0,27	0,2	0,11	0,22	0,06	0,25	0,18	-0,09	1	-0,14
Kasselyd	-0,91	-0,92	-0,86	-0,97	-0,97	-0,98	-0,96	-0,96	-0,97	-0,89	-0,95	-0,95	-0,96	-0,85	-0,56	-0,14	1

Fig. VI: Correlation matrix of the sensory descriptors in Test 2. Based on average values across all musical excerpts, repetitions and listeners. The Pearson correlation formula were used for the calculations. The sensory descriptors are ordered in accordance with their correlation with Basstyrke. Red cells have a correlation coefficient $|r| \geq 0.87$.

one dimension explains the majority of the variability in the data despite the large number of sensory descriptors (Notice, that the ratings for the sensory descriptor "Kasselyd" have been mirrored to ease the interpretation of the plot).

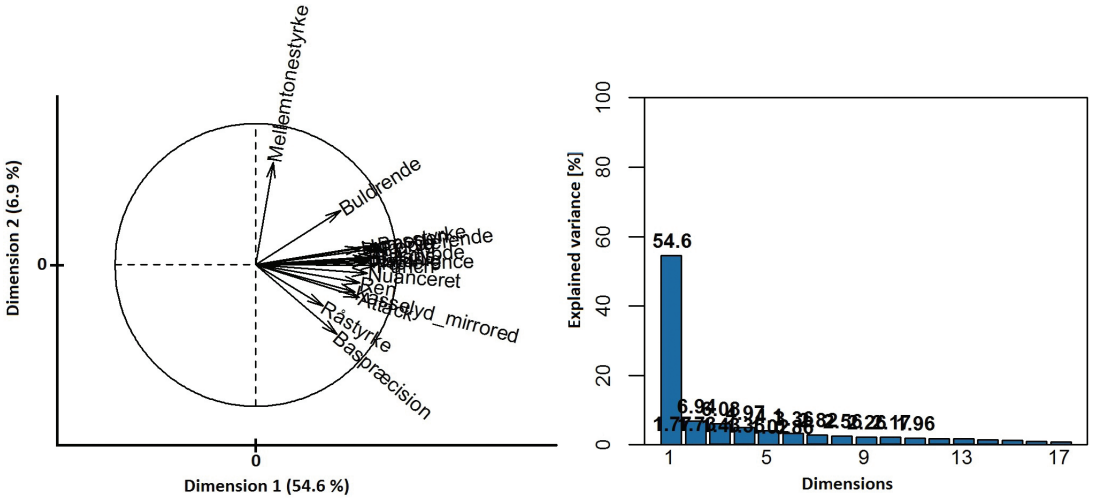


Fig. VII: Principle compoment analysis of the sensory descriptors in Test 2. Based on average values across the PC and SC musical excerpts, all repetitions, and all listeners. *Left:* Loadings plot with Dim. 1 vs. Dim. 2. *Right:* Scree plot.

4.2 Discussion 2

Three sensory descriptors stood out from the correlation matrix in Fig. VI: Mellemtonestyrke, Buldrende and Råstyrke. With regards to Mellemtonestyrke (midrange strength), the systems were either very similar or the midrange was used as an reference for evaluation of bass and treble strength and thus considered to have a "neutral" strength. Another plausible explanation was expressed by listeners, whom viewed the frequency range of the sensory descriptor as too large for evaluation as a whole, making it a difficult task to evaluate overall. Consequently, its position as separated from the main bulk of sensory descriptors in the PCA plot, may be caused by noisy evaluations rather that a real difference in perception.

With regards to Råstyrke, the method of evaluation was different for this sensory descriptor, which might explain the lack of correlation with the other sensory descriptors, but it could also be due to a real difference in perception between this sensory descriptor and the others. The short vector in the PCA plots suggested variance explained by other dimensions than 1 and 2, but

the low explained variance in the remaining dimensions, suggests a that the remaining variance might be explained by noise in the data.

Several potential reasons for the overall high correlations were considered:

1. The systems' perform consistently across all sensory descriptors, i.e. a "good" system is good on all parameters and a "bad" system is bad on all parameters.
2. The listeners had a poor understanding of the sensory descriptors and thus gave a noisy, neutral or hedonic evaluation of the systems.
3. One or more bias' in the test design caused the large correlation.

Based on the large general experience with perceptual product characterisation in SenseLab and on the informal listening by experienced listeners, the first reason listed above was not believed to be the main reason. Consequently the investigation focus were on the two remaining potential reasons. They are discussed in the next two subsections.

Test design bias investigation

One concern in regards to the test design, was that the "baffle" setup caused a bass boost due to the influence from the back wall and the interaction between the back wall and the baffle. The effect being characteristics shared by all systems, due to the general influence from the environment being larger than the contribution from the loudspeaker differences. The effect of the back wall was not initial believed to constitute a problem as the systems and not the loudspeakers were evaluated. Investigation of the measurements revealed that the room influence may have acted as a confounding factor, similar to loudness [11, 12], i.e. the domination of bass strength differences may affect the evaluation of the remaining sensory descriptors significantly. If this is the main cause of the high correlations within sensory descriptors, it does, however, imply that not only sensory descriptors related to timbre are affected by the domination of bass.

In Fig. VIII the frequency responses of the loudspeakers in Test 2 are depicted, normalised around the mean value in each 1/3rd octave band. The figure depicts a boxplot for every 1/3-octave band scaled by an estimation of the JND within each band [13]. This processing technique allowed for a easy visual overview of degree of system differences in a frequency resolution, which approximates that of the human auditory filters.

From the figure it is clear that the differences at mid- and high frequencies above $\approx 800\,Hz$ are small compared to the large differences below; Especially in the three 1/3-octave bands with centre frequencies of 36.5 Hz, 46.0 Hz, and

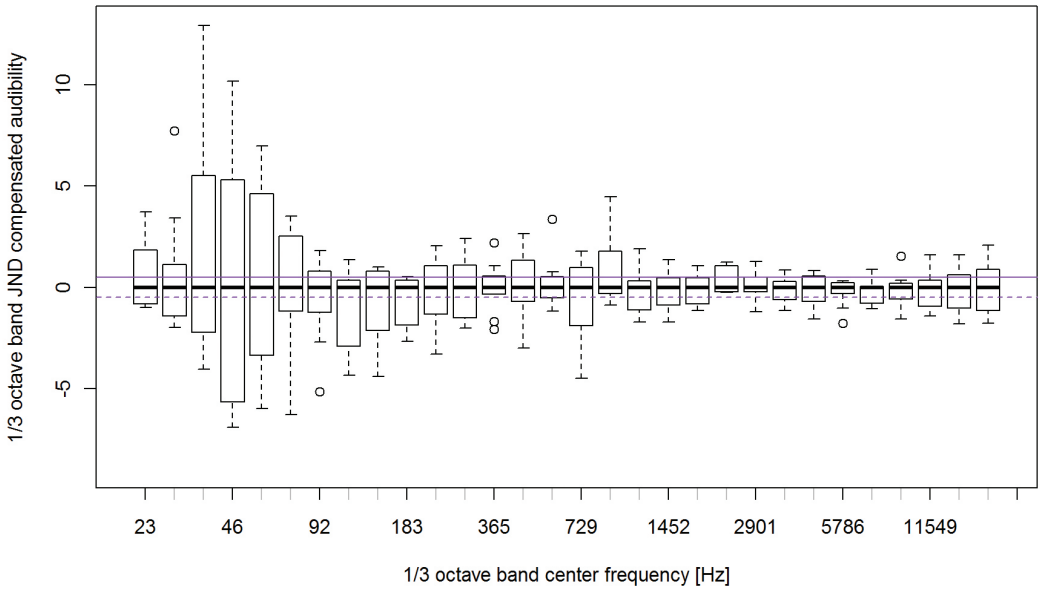


Fig. VIII: *Test 2*: Boxplot of the frequency response of the systems in Test 2 calculated for each 1/3rd octave band. Each boxplot is individually scaled with the corresponding JND of the 1/3rd octave band and the vertical axis, accordingly, has a scale with a JND unit. The dotted purple lines mark the $\pm 1\,JND$ limits. The whiskers of the box spans up to 1.5 times the heigth of the box, but never further than the most extreme data point.

57.9 Hz, i.e. the area of the cut-off frequency of a typical loudspeaker. The JNDs of narrow band noise are however not a precise indication of whether differences are audible in a full frequency spectrum with masking effects and more. Aditionally JNDs are not well established for low frequencies neither in [13] or elsewhere; For frequencies below 250 Hz the JND for 250 Hz was used in the scaling in Fig. VIII.

Comparing with the frequency responses of the systems in Test 2 with the systems in Test 1 (see Fig. IX) the bass domination was a less dominating factor in Test 2.

Another concern, regarding a potential test design bias, was the localisation

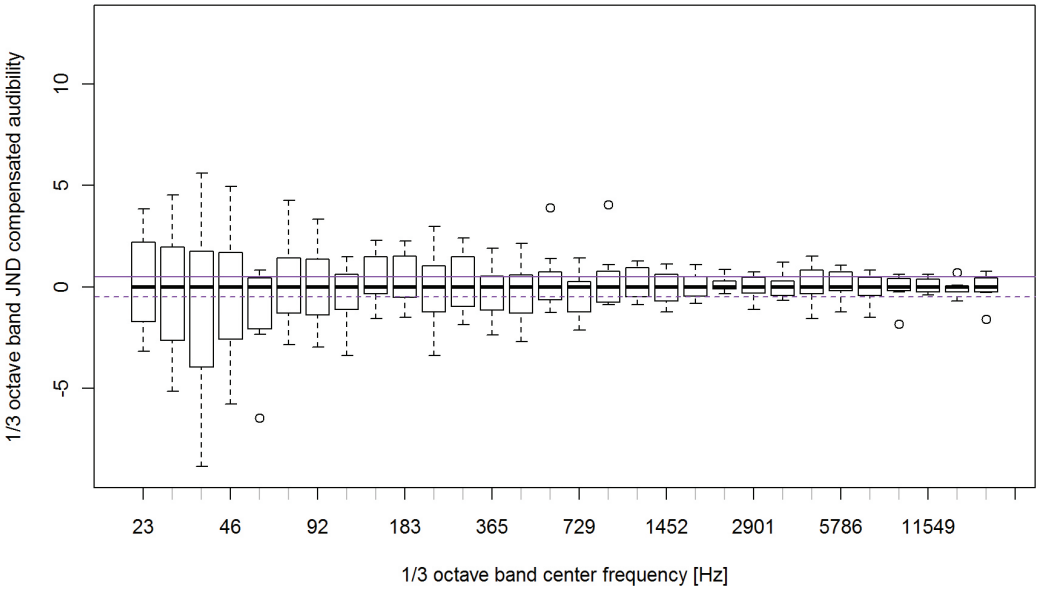


Fig. IX: *Test 1*: Boxplot of the frequency response of the systems in Test 1 calculated for each 1/3rd octave band and normalised around the mean value. Each boxplot is individually scaled with the corresponding JND of the 1/3rd octave band and the vertical axis accordingly has a scale with a JND unit. The dotted purple lines mark the ± 1 JND limits. The whiskers of the box spans up to 1.5 times the heigth of the box, but never further than the most extreme data point.

cues caused by the positioning of the loudspeakers. If a system could be identified based on localisation, the system may have been linked by the listener to a hedonic quality, such as "bad" or "good", potentially affecting the ratings of all sensory descriptors. In Test 1, where most systems were quite similar, this may only have been a problem for the low performing System 2, with the Homebuild loudspeaker, but for the more different systems in Test 2, the effect could be larger. If the listeners did not have a clear concept of

the sensory descriptors under evaluation, the risk of a hedonic bias is likely to become larger. The listener performance-concerns are discussed in the following section.

Listener performance

The uncertainty of whether lacking listener performance were a significant factor in the highly correlating sensory descriptor ratings, was investigated by using the eGauge metrics [14] developed in a collaboration between DELTA SenseLab and Nokia. The metrics consists of Discrimination, Reliability, and Agreement. For each metric a statistical noise-floor can be calculated indicating a performance level, which the listeners should be above to be beneficial for inclusion in the statistical analysis of the test results. While a listener must be above the noise-floor for Discrimination and Reliability to perform well, the Agreement metric is more complex to interpret. One listener with a great understanding could be in disagreement with the average panel ratings, if the average panel has a poor understanding of the sensory descriptor. The one good listener is, however, still diluting the data from a statistical point of view, as he or she would constitute an outlier. Concentrating on Reliability and Discrimination, all listeners were above the noise-floor for four of 16 sensory descriptors: Basstyrke, Fyldig, Nærværende and Naturlig. For an additional five sensory descriptors, one listener was below the noise-floor for either Discrimination or Reliability. For six sensory descriptors two or more listeners were below the noise-floor for Agreement, suggesting that the sensory descriptor was not well understood. The effect of a few listeners with poor performance becomes problematic due to the limited number of trained listeners in the specialized expert assessor panel (SEAP).

An example of a problematic panel performance is depicted in Fig. X[6] for the sensory descriptor Råstyrke. Here, most listeners perform well with regards to Discrimination and Reliability, i.e. are above the noise-floor of both metrics in the top-right square. The majority of the listeners are however below the Agreement noise-floor, indicating that listeners have very different criteria for evaluation. This suggests that the special method of evaluation for this sensory descriptor may have been too unrestricted to lead to homogeneous results.

The lacking listener performance may not conclusively explain the high correlations between sensory descriptor ratings, but was nonetheless an unwanted weakness and handled by starting up weekly training of the listeners

[6]Notice that only seven listeners are depicted in Fig. X. This occurs if an listener rates a system with the exact same score in all repetitions, which lead to an eGauge value of infinity. This is a limitation of the eGauge method in its current version.

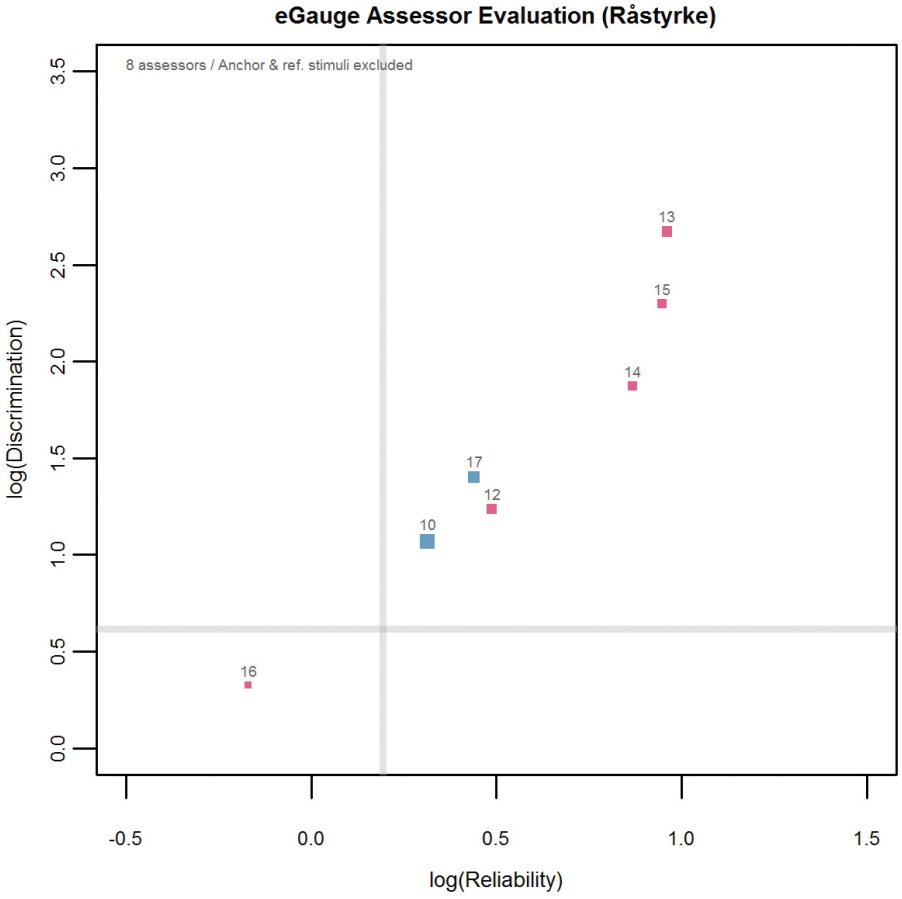


Fig. X: eGauge overview plot for the sensory descriptor Råstyrke in Test 2. Each coloured square represents one listener's score. The grey lines representes the noise-floor for Discrimination and Reliability respectively. A blue colour indicates that the listener is above the Agreement noise-floor and a red that the listener is below the noise-floor. The size of the square indicates the distance to the Agreement noise-floor limit, i.e. a large square is far above or below the noise-floor. A good listener is positioned in the top right corner and (normally) represented with a large blue square.

in SEAP. The training prior to the two listening tests had consisted of including them in the sensory descriptor elicitation process as well as six hours of unsupervised home training. This practise was changed to supervised training to gain a better impression of their understanding of the sensory descriptors, which could further illuminate whether lack of training was the cause of the problems.

5 Electro-acoustical measurements

Electro-acoustical measurements were made of the 15 systems with the purpose of modelling the relation between these and the perceptual evaluations. The measurements were made at the listening position as previously mentioned. They were performed separately on the Test 1 "baffle" of systems and the Test 2 "baffle" of systems with each loudspeaker positioned as it were during the respective listening test.

Prior to the measurements, the overall sound level of systems were equalised and calibrated using a band-limited (80 Hz to 14 kHz) pink noise test-signal and measured with a Sound level meter. The A-weighted loudness equalization of the listening tests were reused, while the overall level was reduced from $70\,dB(A)$ to $67\,dB(A)$. The reduction in level was needed to avoid damaging the smaller loudspeakers, which distorted at the listening test level when playing the logarithmic sweep. The measurement equipment is presented in Table IV.

Product	Manufacturer	Model
Sound card (ADAT)	RME	Fireface 800
Digital to Analogue converter	RME	M32-DA
Microphone (free-field)	Brüel & Kjær	4190
Dual Microphone Power Supply	Brüel & Kjær	5935
Sound level meter	NTi	XL2
Head-and-Torso-simulator (HATS)	Brüel & Kjær	4100D
Power amplifier (2x80W)	NAD	C356BEE
PC software (measurements)	Listen Inc.	SoundCheck 12

Table IV: Electro-acoustic measurement equipment

5.1 Capturing the sound field in the listening position

The measurement setup consisted of six measurement points chosen to sample the sound field surrounding the position of an average listener's head.

Fig. XI depicts these measurement points. The measurement points were positioned with a 60° angle between each. Three positions were 107 cm above the floor and three were positioned 113 cm above the floor, with every second point being positioned in the same vertical plane. These heights were chosen to be centred around 110 cm, which is commonly referred to as the average listening height (approximate height of the ear canal opening above the floor). The position pattern is not aimed at taking listener movement during a listening test into account, but rather account for differences in heights and sitting positions of the listeners. The six measurements points naturally stabilises the measurements in the mid- and high frequency range, due to the spatial averaging over the 6 points which were asymmetrically positioned with respect to the sound sources.

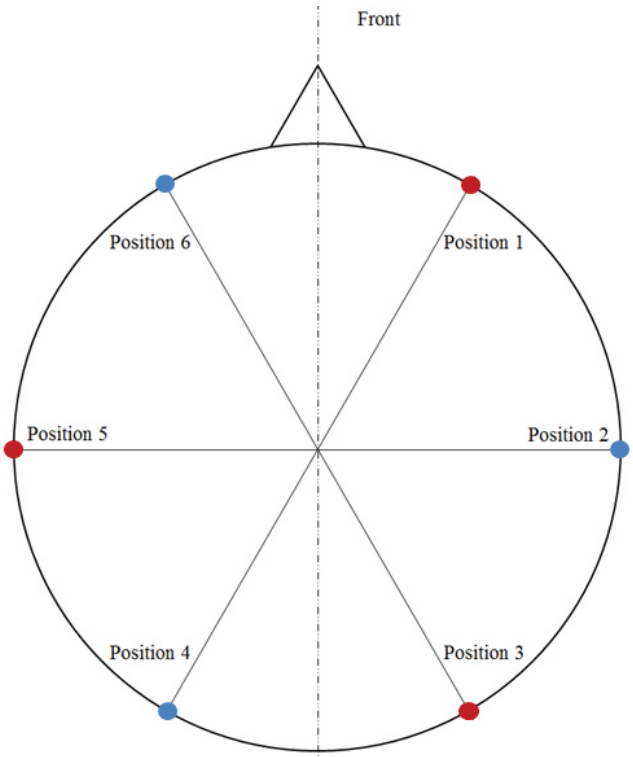


Fig. XI: Measurement positions. Position 1, 3, & 5 were 107 cm above the floor and 2,4, & 6 were 113 cm above the floor. The diameter of the circle was 25 cm.

5.2 Type of measurements

Frequency responses were measurements using SoundCheck. Both a logarithmic sine sweep and a pink noise test-signal were measured for comparison. The logarithmic sweep has the advantage of being deterministic, fast to measure, and providing a high signal-to-noise (SNR). It was measured using a time selective response (TSR) method (also referred to as a time-gated measurement), where the TSR window was set to $\pm 500\,ms$; thereby using 500 ms on each side of the calculated impulse response to calculate the frequency response. The long window corresponds to measurements of frequency content down to 1 Hz. Reducing the window size would decrease the influence of noise, as noise will be recorded as well during the time window. Consequently, it would also remove the influence from reverberations, which contributes to the perceived systems. Below $63\,Hz$, the reverberation time of the listening room exceeds 500 ms, and thus the measured frequency content of the systems slightly underestimates the level of the very low bass.

The pink noise test-signal takes longer to measure, but is less affected by background noise due to the longer time averaging. The length of the noise stimuli determines the uncertainty of the measurement due to the random nature of the signal. For these measurements a length of 20 seconds was used and a 1/6th octave band analysis. A 1/6th octave band resolution was chosen, believed to be better than the human auditory system in the audible frequency range. A 20 second pink noise measurement was analysed using a 1/6th octave band resolutions leading to a theoretical uncertainty [15] (Eq. 3.9.a, p. 96) at $20\,Hz$ for the individual measurement of:

$$\begin{aligned} f_c &= 20\,Hz \\ B &= f_c * 2^{1/12} - f_c / 2^{1/12} = 2.3\,Hz \\ \epsilon &= \frac{4.34}{\sqrt{T_A \cdot B}} \\ \epsilon &= \frac{4.34}{\sqrt{20\,s \cdot 2.3\,Hz}} \\ \epsilon &= 0.64\,dB \end{aligned} \tag{I}$$

Where ϵ is the estimated standard deviation of the assumed normal distribution of the pink noise signal in 1/3-octave band resolution, i.e. there is a 95.5% chance of the true value being within $\pm 2\epsilon = \pm 1.28\,dB$. The uncertainty decreases at higher frequencies, as is evident from the above equation. The source uncertainty decreases further with the averaging over positions to a sixth of the calculated ϵ, i.e. to $\epsilon = 0.11\,dB$, at which point the uncertainties of the calibration, the measurement equipment, positioning of the microphones etc. are larger.

5.3 Frequency response overview

In Fig. XII the differences in frequency responses among the 15 systems are plotted. The depicted frequency responses are based on the measurements using the pink noise test-signal. Clearly audible differences are found up to approximately 400 Hz, which coincides with the definition DELTA SenseLab have for the upper limit of bass [16]. While the remaining frequency range have limited area of the boxes within the $\pm 1\, JND$ dotted line, it is worth remembering that only 50% of the data points lies within the area of the box.

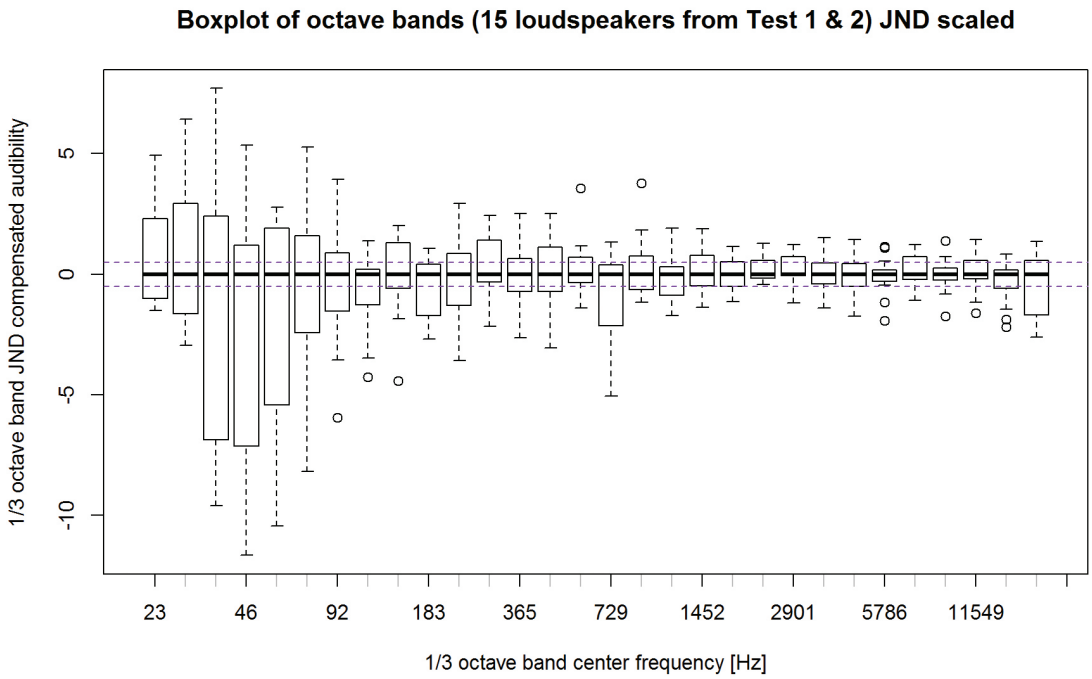


Fig. XII: Boxplot of the frequency response of the systems in both Test 1 and Test 2 calculated for each 1/3rd octave band and normalised around the mean value. Each boxplot is individually scaled with the corresponding JND of the 1/3rd octave band and the vertical axis accordingly has a scale with a JND unit. The dotted purple lines mark the $\pm 1\, JND$ limits. The whiskers of the box spans up to 1.5 times the heigth of the box, but never further than the most extreme data point.

5.4 Alternative measures

In addition to the electro-acoustic measurements presented in Section 5.2 recordings of the stimuli was also captured in the listening position with a B&K HATS and a single microphone respectively, using Adobe Audition 3.0 and the same hardware used for the electro-acoustical measurements. The main purpose was to obtain recordings, which could be used for a potential comparison of perceptual evaluations in-situ vs. perceptual evaluations using headphone auralization.

6 Concluding remarks

The idea behind the "baffle" setup was to obtain an intermediate method of capturing perceptual and electro-acoustical measurements of sound events generated by loudspeakers - audio reproduction sources. The goal was to collect data for predictive modelling of perceptual characteristics, gain information about the suitability of the elicited sensory descriptors, and continue the training of the listeners in SEAP.

Unfortunately the results of the perceptual evaluations showed severe problems in the data quality as described in this appendix. The potential causes of the problems are discussed, but the main reason was not conclusively identified. Since this was an intermediate step on the path to being able to perform in-situ perceptual evaluations of loudspeakers - not "systems" - no further measures were taken to untangle the root of the problem. The transition from unsupervised training to supervised training and the increase in training intensity was, however, undertaken to maximize the likelihood of obtaining better results in future loudspeaker listening tests.

With regards to the measurement technique with six positions, further investigation of how to capture a sound field representative of the listening position was undertaken (as described in Section 6).

Acknowledgement

The listening tests were designed in a collaboration between SenseLabs Test leader (Tore Stegenborg-Andersen), the Ph.D. student of this thesis and the company supervisor of the Ph.D. project (Torben H. Pedersen). The supervised training of the listeners in SEAP were also initiated and led by Torben H. Perdersen. All electro-acoustical measurements were performed by the Ph.D. student with SoundCheck assistance from the Test leader. The analyses of this appendix were discussed, expanded on and improved by input from Søren Bech, Flemming Christensen, and Torben H. Pedersen. The listening test software was programmed by Søren V. Legarth.

References

[1] C. P. Volk, S. Bech, T. H. Pedersen, and F. Christensen, "Five aspects of maximizing objectivity from perceptual evaluations of loudspeakers: A literature study," in *AES Convention 138*, Audio Engineering Society, 05 2015. Convention Paper 9230.

[2] T. H. Pedersen and N. Zacharov, "The development of a sound wheel for reproduced sound," in *AES Convention 138*, Audio Engineering Society, 05 2015. Convention Paper 9310.

[3] Shirley Caesar, "Stand Still Feat. John P. Kee [Musical track]." *Album: Stand Still*; Word, Inc., United States, 1993.

[4] Paula Cole, "Tiger [Musical track]." *Album: This Fire*; Warner Bros., United States, 1996.

[5] Eliane Elias, "Chega De Saudade [Musical track]." *Album: Bossa Nova Stories*; Blue Note Records, United States, 2003.

[6] T. H. Pedersen, "Perceptual characteristics of audio: The sound wheel can be used to provide objective description of the sound," Technical document Tech document no. 7 2015, DELTA SenseLab, 03 2015.

[7] O. Tomic and H. Risvik, "PanelCheck (1.4.2) [Computer software]." Nofima Mat & Technical University of Denmark & University of Copenhagen (Department of food science), `http://www.panelcheck.com`, Denmark & Norway, 2012.

[8] ITU-R, "Method for the subjective assessment of intermediate quality level of coding systems," Recommendation ITU-R BS.1534-1, International Telecommunication Union Radiocommunication Assembly (ITU-R), 01 2003. MUSHRA.

[9] N. Schinkel-Bielefeld and A. K. Leschanowsky, "How much is the use of a rating scale by a listener influenced by anchors and by the listener's experience?," in *Audio Engineering Society Convention 138*, 05 2015.

[10] E. C. Poulton, *Bias in quantifying judgements*. Hove, East Sussex, U.K.: Lawrence Erlbaum Associates Ltd., 1989. ISBN: 978-0-86377-105-7.

[11] A. Gabrielsson, "Perceived sound quality of reproductions with different frequency responses and sound levels," *The Journal of the Acoustical Society of America*, vol. 88, no. 3, p. 1359, 1990.

[12] A. Illényi and P. Korpássy, "Correlation between loudness and quality of stereophonic loudspeakers," *Acta Acustica united with Acustica*, vol. 49, pp. 334–345, 12 1981.

[13] M. Florentine, "Level discrimination as a function of level for tones from 0.25 to 16 kHz," *The Journal of the Acoustical Society of America*, vol. 81, no. 5, pp. 1528–1541, 1987.

[14] G. Lorho, G. Le Ray, and N. Zacharov, "eGauge—a measure of assessor expertise in audio quality evaluations," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, 06 2010.

[15] R. B. Randall, *Frequency analysis*. Nærum, Denmark: Brüel & Kjær, 3. ed., rev., 1. print ed., 1987. ISBN: 978-87-87355-07-0.

[16] T. H. Pedersen, "The semantic space of sounds - lexicon of sound-describing words," Lexicon ISBN 978-87-7716-036-3, DELTA SenseLab, 05 2008. Revised edition.

SUMMARY

Perception of the reproduction characteristics of headphones and loudspeakers are presently not directly related to the many technical measurements made in the physical domain. This have made it difficult to interpret traditional measurements of headphones and loudspeakers in terms of how physics affects perception.

In this project a number of audio metrics are presented, which describes perceptual characteristics in terms of properties of the physical acoustical output of headphones and loudspeakers. The audio metrics relies on perceptual models for estimations of the how these acoustical outputs are processed in the human auditory system.

The work was carried out in the period September 2013 to September 2016 and was a collaboration between Aalborg University and DELTA SenseLab with funding from the Danish Agency for Science, Technology and Innovation.

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-825-3

AALBORG **UNIVERSITY PRESS**