

FP7-ICT-2013-C TWO!EARS Project 618075

Deliverable 6.2.3

QoE model software, final version



WP6 *

November 30, 2016

* The TWO!EARS project (<http://www.twoears.eu>) has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 618075.

Project acronym: TWO!EARS
Project full title: Reading the world with TWO!EARS

Work package: 6
Document number: D 6.2.3
Document title: QoE model software, final version
Version: 1

Delivery date: 30. November 2016
Actual publication date: 01. December 2016
Dissemination level: Public
Nature: Report

Editor(s)/lead beneficiary: Alexander Raake
Author(s): Alexander Raake, Hagen Wierstorf, Fiete Winter, Sascha Spors, Chung Eun Kim, Armin Kohlrausch, Thomas Walther, Jens Blauert, Tobias May, Patrick Danès
Reviewer(s): Jonas Braasch, Dorothea Kolossa, Bruno Gas, Klaus Obermayer

Contents

| | | |
|----------|--|----------|
| 1 | Executive summary | 1 |
| 2 | Activities during year three | 3 |
| 2.1 | Introduction | 3 |
| 2.2 | How can <i>Quality of Experience</i> for spatial audio systems be assessed in listening tests? | 3 |
| 3 | Listening tests on sound quality | 7 |
| 3.1 | Comparing different spatial audio reproduction systems | 8 |
| 3.1.1 | Methods | 10 |
| 3.1.2 | Results | 17 |
| 3.1.3 | Discussion | 20 |
| 3.2 | Finding the sweet spot in 5.0 surround | 21 |
| 3.2.1 | Method | 21 |
| 3.2.2 | Results | 23 |
| 3.2.3 | Discussion | 23 |
| 3.3 | Combined PC and MDS approach for listening preference | 25 |
| 3.3.1 | Methods | 25 |
| 3.3.2 | Results | 26 |
| 3.3.3 | Discussion | 29 |
| 3.4 | Scene-specific quality assessment: Influence of source location and colouration combined | 30 |
| 3.4.1 | Experimental design and procedure | 30 |
| 3.4.2 | Results and analyses | 33 |
| 3.4.3 | Effect of colouration | 34 |
| 3.4.4 | Effect of source location | 35 |
| 3.4.5 | Discussion | 35 |
| 3.5 | Concert hall recordings – preference and perceptual attributes | 36 |
| 3.5.1 | Experimental design and procedure | 37 |
| 3.5.2 | Results and analyses | 38 |
| 3.5.3 | Discussion | 44 |
| 3.6 | Compression of interaural level differences | 44 |
| 3.6.1 | Test method and procedure | 45 |
| 3.6.2 | Results and analyses | 46 |

| | | |
|----------|---|-----------|
| 3.6.3 | Discussion | 51 |
| 4 | Final quality model | 53 |
| 4.1 | Implementation of the van Dorp Schuitman (2011) model within the TWO!EARS AFE framework | 53 |
| 4.1.1 | Model structure | 53 |
| 4.1.2 | Implementation of model | 55 |
| 4.1.3 | Discussion and outlook | 57 |
| 4.2 | Preference prediction | 57 |
| 4.2.1 | Motivation and modeling goal | 57 |
| 4.2.2 | Perceptual data | 58 |
| 4.2.3 | Modeling approach | 62 |
| 4.2.4 | Feature extraction | 63 |
| 4.2.5 | Model variations using different feature subsets | 65 |
| 4.2.6 | Model evaluation | 65 |
| 4.2.7 | Modeling performance | 69 |
| 4.2.8 | Discussion | 76 |
| 5 | Conclusions | 79 |
| | Bibliography | 81 |

1 Executive summary

The TWO!EARS model will be evaluated in the context of two possible applications. These are *Dynamic Auditory Scene Analysis* (DASA) and *Quality of Experience* (QoE). The first application is discussed in D 6.1.3. The present document focusses on the work towards the QoE application of the model.

For the *Quality of Experience* assessment, the work focussed on model development, and very important, on acquiring ground-truth data in a number of listening tests. For the sound quality and QoE assessment of high-quality spatial audio systems as planned in TWO!EARS, there were no results and sound data available to the consortium prior to the project that could be used for model development. Since quality perception occurs in the mind of the listener and existing models did not support the types of systems addressed in TWO!EARS, no automatic quality-labelling of scenes was available to support the ground-truth data collection. In D 6.2.1, we described a number of test methods and a first set of trial tests. In D 6.2.2, the testing and modelling work for the individual features localization accuracy and coloration were reported.

The work in year 3 focused on running tests and developing the sound quality / *Quality of Experience* modelling approaches for a proof-of-concept of the TWO!EARS framework for this application. A variety of test paradigms was used to address different properties of the TWO!EARS modelling framework, such as scene-specific evaluation, combination of features to sound quality and *Quality of Experience*, and verification of quality-specific model implementations.

In the first chapter, we will discuss the general challenges of QoE models and the parallel work streams followed in year 3 of TWO!EARS. It turns out that for the considered high-quality systems, the modeling is tightly dependent on the actual listening experiments. For the *Quality of Experience* model proof, we have picked a specific series of tests to show the general way of modeling.

All test data collected during the final year is contributed to the public TWO!EARS database, which is described in Deliverable D 1.3. All the modeling approaches are published with the final version of the model.

Since both the quality-feature models (D 6.2.2) and, during year 3, also the preference prediction modelling appeared to be realizable based on the TWO!EARS framework, a proof

1 Executive summary

of concept for the sound quality and *Quality of Experience* domain could successfully be delivered.

2 Activities during year three

2.1 Introduction

The TWO!EARS project aimed at developing an intelligent, active computational model of auditory perception and experience that operates in a multi-modal context. Evaluating *Quality of Experience* (QoE) for spatial audio systems is one of the two proof-of-concept applications of the model.

The present report summarises the quality model related activities and developments during the third year until the end of the project. It builds on the *Quality of Experience test method specification* provided in Deliverable D 6.2.1, and provides a wide range of new results from listening tests. In addition, it extends the model approach presented in Deliverable D 6.2.2 and predicts single low-level and high-level attributes in order to judge the perceived sound quality of spatial audio reproduction systems.

This document is structured as follows: In the subsequent sections of this chapter, we explain the insights we have gained during our work on assessing QoE in listening tests and modelling the results. We will discuss the challenges of investigating QoE for spatial audio systems.

In Chapter 3, we present the results from the listening tests carried out during the third year. Some of the data will then be modelled in Chapter 4.

2.2 How can *Quality of Experience* for spatial audio systems be assessed in listening tests?

In the following, the ambitious goals set for the *Quality of Experience* (QoE) proof-of-concept domain will be discussed in the light of the achievements reached for this area by the end of year 3.

TWO!EARS addresses spatial audio that is presented via loudspeakers or headphones. All individual elements of the end-to-end chain from sound creation to presentation may play a role for their perception (Spors *et al.*, 2013). The project addresses the performance

of spatial audio systems in terms of *sound quality* and QoE (see for example (Raake, 2016)).

Sound quality explicitly refers to how the influence of the technical system is perceived and evaluated. It is sometimes referred to as *Basic Audio Quality*, corresponding to the terminology used in standards such as MUSHRA, BS.1534 (ITU-R BS.1534). Sound quality evaluation for loudspeaker-based systems may follow a “spatial and timbral fidelity” paradigm (Rumsey *et al.*, 2008). It relates to findings initially obtained for stereophonic systems that the variance in sound quality tests is explained to 70% by “timbral fidelity” and 30% by “spatial fidelity” (Rumsey *et al.*, 2005).

Assessing QoE relates to the audio experience in a more holistic manner, and implies that the listener may not be explicitly focused on the involved technology or its assessment. Examples for attempts to assess QoE for audio systems can be found in Schoeffler and Herre (2016), Lepa *et al.* (2013), Wilson and Fazenda (2016a). Due to the general difficulty of QoE assessment, most of the literature from the audio technology domain is on sound quality. The challenges related with sound quality and QoE evaluation for spatial audio have been listed in Deliverable D 6.2.2 and are summarized here so as to explain the subsequent steps of the project activities in Task 6.2 of WP6 during year 3:

- Sound quality and QoE for systems such as Wave Field Synthesis and Higher Order Ambisonics is difficult to assess in tests or by means of models. This is due to the fact that quality-differences for such systems are typically small as compared to those encountered for example with different low-bitrate audio coding algorithms.
- The perceptual reference for spatial sound quality evaluation is much less well established than for other audio systems such as codecs. Often times, especially with popular music, no explicit reference may be available at all, since there typically is a strong interaction between the music creation, recording, postproduction and presentation stages. Here, listeners may make evaluations in comparison to internal references that correspond to the better established stereophonic mixing and processing chain as it is used for 2-channel or possibly 5.1 stereo.
- Listening tests are the primary approach for sound quality evaluation and serve for generating the ground-truth data later used during model development¹. In some cases, it is unknown in prior whether the assumed effects will actually be observed in

¹ Since none of the project partners was able to get access to available audio stimuli and respective test data from existing campaigns such as those run in MPEG audio, all ground-truth data used in Two!EARS modelling had to be created by the project itself. In contrast to the Dynamic Auditory Scene Analysis application area, stimuli annotations in terms of sound quality or QoE can only be obtained by tests with human subjects, also since no existing spatial-audio-quality models are available to assist by automatic annotation of stimulus databases.

a given test. Hence, the test results have to be awaited to know whether they can be used as ground-truth data for modelling.

- The link between auditory features and a certain level of QoE for listeners may be difficult to establish with a rather limited set of data.

Because of these challenges and the results obtained during the QoE assessment campaigns during the early stages of TWO!EARS, it was decided to take a pragmatic approach suited for proof-of-concept, namely to address sound quality in terms of the two separate features *colouration* and *localisation*, and to work towards direct sound quality and QoE assessment using paired comparison (PC) tests.

Our proof-of-concept results for modelling individual features with the Two!Ears model framework are described in D 6.2.2. The present document primarily addresses the tests and modelling campaigns related with the assessment of sound quality and QoE. Using a PC-type preference test paradigm was motivated by recent findings, which showed that asking for *video quality* or *Basic Audio Quality* can introduce a bias in the ratings towards signal clarity or timbral features, observed for both audio and video quality tests (Zacharov *et al.*, 2016, Benoit *et al.*, 2008, Lebreton *et al.*, 2013).

PC-type preference tests can avoid such biases and simultaneously address the two first challenges stated above. In addition, it can be assumed that in a PC-type preference test under laboratory conditions – and with the correct choice of stimuli – the resulting judgment is closer to an assessment of actual QoE, as the test subjects are asked to rate which presentation they prefer, instead of rating sound quality.

Early in year three, the final approach for the QoE proof-of-concept was determined. Here, it was decided to primarily focus the available resources in this area on the proofing of major functionalities of the TWO!EARS framework. Based on the plans laid out in D 6.2.1 and D 6.2.2, four different parallel work streams were prioritised during year three, evaluating which of these would be most suited for sound quality- and QoE-related modelling:

1. **Comparing different spatial audio reproduction systems:** Listening tests and model addressing the comparison of the different spatial reproduction systems stereo, surround, and WFS. This includes a scene-based paradigm with ecologically valid stimuli, and development of a scene-based model using the TWO!EARS framework. To enable this it was planned to extend the reproduction-specific mixing of object-based audio tracks to create realistic, clearly audible and object-specific effects.
2. **Finding the sweet spot for 5.1 surround:** Test and model for active selection of the preferred listening position in a given context of 5.1 stereo as reproduction method for a musical piece recorded with different microphone set-ups. Here, a series of tests was planned, having subjects chose the preferred position when visual information

about the loudspeaker set-up was provided, and when this was not the case. This way, the impact of visual information and respective adaptation of decisions was planned to be addressed, alongside with the evaluation of features that result in a preference for a given position.

3. **Combined PC and MDS approach for listening preference:** Work on combined PC- and MDS-based tests for a number of spatial audio systems. In case of successful completion, modelling campaign should be started to predict the resulting dimensions based on features from TWO!EARS' set of knowledge sources (KSs) including lower-level features from the Auditory Front End (AFE).
4. **Scene-specific quality assessment:** Extend the work on simpler scenes using the initial scene from Raake *et al.* (2014) with further colouration and localization related effects. It was planned to also evaluate the sound quality in PC tests and establish a connection to the attributes colouration and localization.

The next Chapter presents results from listening tests for all four work streams. On the basis of those listening test the work on actual modeling of sound quality and QoE is presented in the Chapter thereafter.

3 Listening tests on sound quality

In the following the experiments carried out during the last year are presented. For most of them, the results, analysis scripts and the stimuli are contributed to the public database described in D 1.3 and on our online documentation (<http://docs.twoears.eu>). For these experiments, a link to the online documentation is provided at the beginning of the corresponding subsection. The publication of the stimuli along with the results has the advantage that for all experiments that applied binaural synthesis the ear signals can directly be fed into a binaural model.

An important goal of the spatial audio quality evaluation work in TWO!EARS has been to reflect the scene- and object-specific quality evaluation performed by human listeners Raake *et al.* (2014), Raake and Blauert (2013). When analysing whether sound quality can be judged and modelled following such a scene-based paradigm, sound quality effects are required that are suited for evaluation according to such an object-specific view. The work presented in Section 3.1 reflects this view, as well as the one presented in Section 3.4, which is a direct extension of our preliminary work with this paradigm presented in Raake *et al.* (2014).

For high-quality systems as addressed in TWO!EARS, the source content used principally is a critical component. For stereophonic systems, this usually is less problematic, since professional audio material is typically mixed and produced with a stereophonic play-out system in mind. The production chain for stereo still mostly follows a classical channel-based paradigm rather than an object-based one, and the underlying raw material is rarely available. For the realistic comparison of different spatial audio systems, as well as the application of the afore-mentioned scene-based evaluation paradigm, we have realised dedicated own spatial audio mixes in the course of the project, see Section 3.1.

For a given spatial audio set-up, the sound quality also depends on the position of the listener. One evaluation paradigm applied in year 3 was to have listeners select the listening position that is connected with the best listening experience – once with visual information about the geometric relation to the loudspeaker set-up, once without (see Section 3.2).

Moreover, we implemented the combined Paired Comparison tests for preference evaluation and multidimensional scaling (MDS) outlined in Deliverable D 6.2.1 to evaluate different spatial audio systems. Here, we used similar settings as for the feature-specific evaluation

described in Deliverable D6.2.2, see Section 3.3.

All those experiments reflect the four parallel work streams introduced in Section 2.2. In addition, further experiments were carried out to support the modelling work. One experiment tested the validity of another binaural model (van Dorp Schuitman *et al.*, 2013) that was identified to be a good candidate for integration into TWO!EARS (see Section 3.5). Another experiment investigated the influence of non-linear inner ear effects on the perception as those are part of the TWO!EARS modelling approach. In this case, the effect of compression on the ILD is investigated in Section 3.6.

3.1 Comparing different spatial audio reproduction systems

🔗 **Published in database:** see online documentation for experiment 1

🔗 **Published in database:** see online documentation for experiment 2

🔗 **Published in database:** see online documentation for experiment 3

This section investigates which of the three different audio reproduction systems 2.1-channel stereophony (stereo), 5.1-channel stereophony (surround), and wave field synthesis (WFS) using a 56.1-channel circular loudspeaker array might be preferred by listeners in the context of popular music. The context of popular music is a very interesting one as it has high practical relevance to most listeners. At the same time, it is a very challenging one as there is no reference or live performance that the different reproduction systems try to match, but an artistic intent or goal that should be reached at the end of the production process of popular music. In practice, this almost always results in track-based recordings of single instruments (e.g., guitar, lead vocal, drums), followed by summing or mixing of these instrument tracks down to a stereo or surround master (Bitzer *et al.*, 2008).

This means there is always the influence of what a mixing engineer is able to achieve with a given reproduction system, not only the physical properties of such a system that determine the listener experience at the end. For the comparison of stereo and surround this influence can be limited by producing a music mix for the surround system and use a down-mixing algorithm to generate the stereo mix. In our opinion, such an approach cannot be meaningfully applied if stereo or surround should be compared with WFS as there are too many differences between those systems. The high number of applied loudspeakers in WFS can lead to stronger problems with colouration due to comb-filter like spectra than in stereo or surround (Wierstorf *et al.*, 2014). Mixing engineers are able to adjust their mixes to such problems, which cannot be achieved by up-mixing algorithms. In addition, stereo and surround are channel-based reproduction systems and in most

cases their content is produced employing a channel-based panning approach. WFS is designed to be independent of the number of applied channels and is in most cases mixed with an object based approach, which can require different techniques and solutions to problems that mixing engineers try to solve in order to convey the artistic intent of a song.

As a result the creation of the musical content for the listening test for the different systems cannot be detached completely from the influence of the mixing engineer. We tried to disentangle this influence by not only varying the reproduction systems, but the critical mixing parameters *EQ*, *positioning*, *reverb*, and *compression* as well.

In the *EQ* process the mixing engineer tries to enhance and correct the frequency content of the mix. In addition, EQing is used besides positioning to unmask content (e.g. separate vocals from other instruments). EQing can have direct impact on the listening preference as bass balance (Wilson and Fazenda, 2015) and brightness (Fung, 1996) can influence the perceived sound quality.

Positioning is another important mixing process as it is one of the most influential tools in creating spaciousness, which is one common goal in the mixing process of popular music. It also ensures that there are no gaps between instruments and that there is a spatial balance between left and right (Pestana and Reiss, 2014). In popular music there exists a well-established pattern for positioning single instruments, like the lead vocals in the center (Mansbridge *et al.*, 2012). A deviation from this pattern might have an influence on listener preference. There are also results reported that show a negative influence of too narrow and too wide arrangements (Choisel and Wickelmaier, 2007).

Compression is known to the audio community from the loudness war, which refers to its usage on the final master of the mixing process. But it can also be used on single tracks in order to make one loud enough compared to another. Some genres, like rock, are not possible without compression on the vocals. Compression is non-linear and has a non-trivial selection of parameters as the effects are not obvious and the parameters are correlated. In addition it can introduce artifacts like pumping, breathing, and low frequency distortion, which in most cases appear if too much compression is used (Giannoulis *et al.*, 2013). This means the influence of compression on listening preference is the one of a best point (Wilson and Fazenda, 2015).

Reverb is in most cases artificial reverb in the context of popular music. It is used to enhance envelopment by providing space and distance information. There seems to be a perfect level between foreground and reverb regarding listening preference (Pestana and Reiss, 2014).

In a first step, a concept was developed that enabled the mixing for the three different reproduction systems with the same goal in mind, but also allowing to achieve the best

that is possible with each system. This concept was then tested for four different songs (experiment 1). As the results of experiment 1 showed that the different songs had no large influence on the results, we continued with only one song, but now changed several mixing parameters in a systematic way (experiment 2). As the goal was to predict listeners' preference ratings at the end by a binaural model, it is desirable to use binaural simulations of the different systems in order to get the input signals for the model. As the binaural simulation is not completely transparent (non-individual HRTFs), we repeated experiment 2 with dynamic binaural synthesis in order to judge the influence the binaural simulation has on the listener preference ratings (experiment 3).

3.1.1 Methods

Apparatus

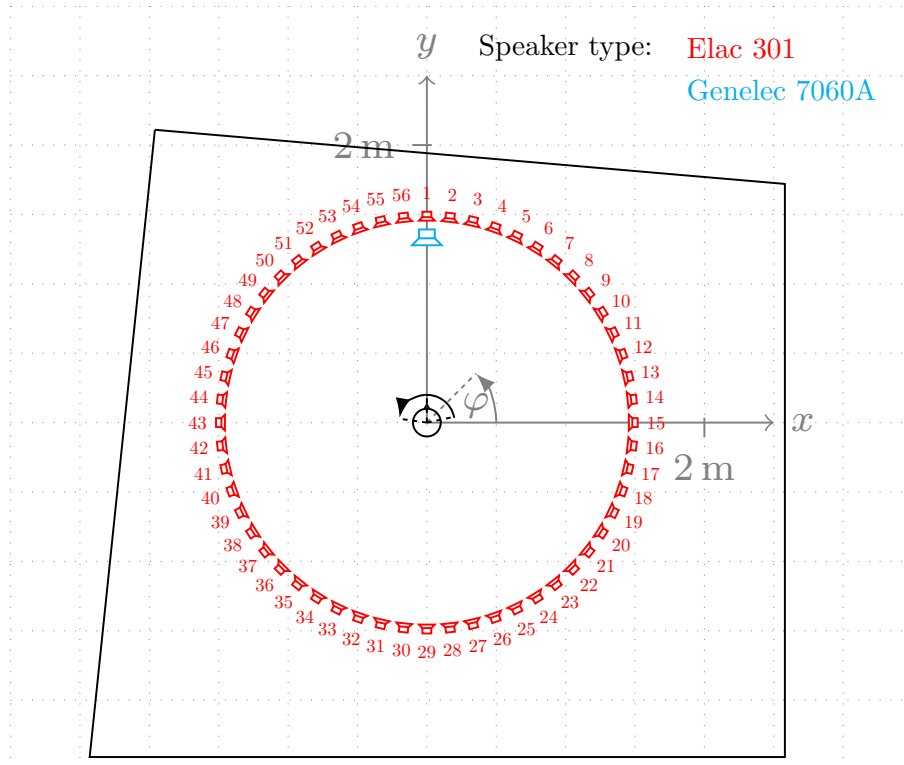


Figure 3.1: Sketch of loudspeaker setup in room Pinta.

Experiment 1 and experiment 2 involved the loudspeaker array installed in an acoustically treated listening room (room Pinta in the Telefunken building of TU Berlin) shown in Fig. 3.1. The listeners sat in a heavy, although rotatable chair in the center of this array,

facing a flat screen, provided with a keyboard. The circular loudspeaker array consists of 56 small loudspeakers (ELAC 301), enhanced by a subwoofer (Genelec 7060A). In a separate room, a computer equipped with a multichannel sound card (RME Hammerfall DSP MADI) played back all sounds. The active powered subwoofer received an analog signal directly from the soundcard's D/A converter. All remaining 56 channels were sent via MADI to the listening room and D/A converted as well as amplified by custom-made units.

Experiment 3 took place in a 83m³ acoustically damped listening room (room Calypso in the Telefunken building of TU Berlin). The listeners sat on a heavy chair wearing open headphones (AKG K601) with an attached head tracker (Polhemus Fastrak). They sat in front of a flat screen placed on a small table and were able to choose between a mouse or keyboard for entering their responses. In a separate room, a computer equipped with a multichannel sound card including D/A converters (RME Hammerfall DSP MADI) played back all sounds. The signals traveled through a head phone amplifier (Behringer Powerplay Pro-XL HA 4700) and analogue cable to the head phones in the listening room, a distance of approximately 5 m.

Audio material and general mixing concept

The audio material consisted of a multitrack recording session with double trackings, mainly for guitars and vocals. It was a moderate tempo pop music piece including deep male vocals, acoustic and electric guitars, bass, drums, shaker and also reverb and delay effects ((unpublished) — Lighthouse). The mixing engineer also recorded the track, ensuring no heavy processing was applied during recording.

For experiment 1, three other stimuli were generated from freely available multi-track recordings. The three recordings consist of a pop-rock song with live feeling and male vocals (The Brew — What I Want¹), presented complete (1) and from its guitar solo bridge (2), a slightly heavier rock song with female vocals (Hop Along — Sister Cities²) and a shorter hip-hop track (Lushlife — Toynbee Suite²) with male rap vocals. In experiment 1, also two versions from the song Lighthouse were presented, again complete (1) and with its very spatial guitar bridge (2).

The mixing process of popular music involves stages that are more or less independent from the involved reproduction system. On the other hand, some stages like panning and reverberation have to be adjusted on each single systems in order to ensure they are used in the best possible way. Applying advanced mixing techniques demands for an adaptation

1 Single tracks downloaded from <http://www.telefunken-elektroakustik.com/download/brew/>

2 Single tracks downloaded from <http://www.medleydb.weebly.com/downloads>

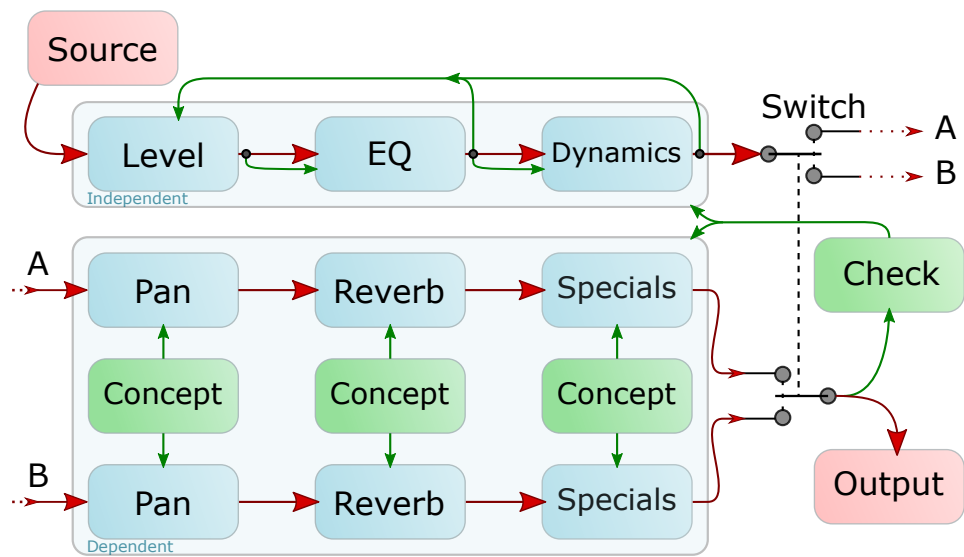


Figure 3.2: Block diagram illustrating the basic multi system sound mixing procedure, from the source material to the finished output. Red corresponds to audio signals, blue to sound processing stages and green shows major dependencies.

to WFS, as most modern mixing techniques are based on channel-based stereophony and not on the model-based approach taken by WFS.

Figure 3.2 shows a block diagram of the basic layout we applied in order to create comparable mixes for different systems. Red arrows symbolize the path of audio signals, small blue boxes the sound processing stages and green shows major dependencies and influences of the mixing engineer. The upper large gray box describes all the system independent steps, the lower one all steps performed on the actual systems.

The system independent mixing took place on an completely independent stereo reference system, very well known by the mixing engineer. There, he applied the optimum amount and type of EQing and compression, also consulting the artists of the song. Small adjustments were applied in the actual WFS system afterwards.

The final system dependent processing included positioning, reverb, and special effects. All system dependent processing was performed directly on the systems involved in the listening experiment. It was guided by an underlying joint concept, which means that every processing and actual mix decisions are consistent across systems. This avoids fundamentally different mix results and ensures comparability. Nevertheless, each system needs idiosyncratic adaptations in order to perform adequately, particularly the handling of sound positioning. Those adaptations, especially the positioning of the single objects in the music scene, followed a conservative handling, which implied avoiding extreme and rather

unconventional settings as well as omitting moving sources. Hence, all main components of a song were positioned in the frontal scene. Lead vocal, snare drum and bass were always positioned in the center. Since further positioning was mainly driven by the available capabilities of each individual system, especially their spatial performance, elements that likely create envelopment were allocated to all directions. In particular, reverbs and delays should exhaust and thereby demonstrate the spatial capabilities.

All special processing, such as modulation based effects, were made as similar as possible between the systems. The final automation and correction/checking stage was again valid for all systems and therefore performed after the switching. The validation affected every stage.

Variation of mixing parameters

For experiment 2 and experiment 3, only the song *Lighthouse* was used, but the mixes for WFS were varied for the following mixing parameter in a systematic way.

EQ Given the reference equalizer settings from the reference mix (WFS), there were alterations in both directions, more and less EQing. Therefore, the amount of boost or cut per filter frequency got scaled. For the condition more EQ (E+), every intervention was exaggerated by doubling the amount of applied filter gain. Respectively, every applied filter gain was halved for condition less EQ (E-). Bypassing every equalizer in each channel finally led to condition no EQ (E--). Note, that this also bypassed all involved high-pass filters.

Positioning Starting from the reference mix (WFS), two different mixes with narrower spaced foreground elements and two mixes with wider spaced foreground elements were produced. The choice of which audio objects belong to the musical foreground is highly content dependent, but was quite obvious in the present music piece. The decision taken for “vocal, drums, guitar” also matches the three most mentioned instruments regarding “like” and “quality” from a previous study by Wilson and Fazenda (2016b). Here, the foreground instrument “drum” was represented by the sound objects “bassdrum” and “snaredrum” and its remaining parts were considered to belong to the background. Besides the lead tracks, common pop music practice includes vocal and guitar (harmony-) double tracks, which were also displaced accordingly.

The reference mix (WFS) represents a very common and modern variant with lead tracks in the center, guitar tracks positioned to the side and double tracks spread symmetrically, compare Fig. 3.3. This arrangement is similar to the stereo mix, however moderately wider. The narrow version (P-) moves all tracks towards the center. The very narrow (P--) mix

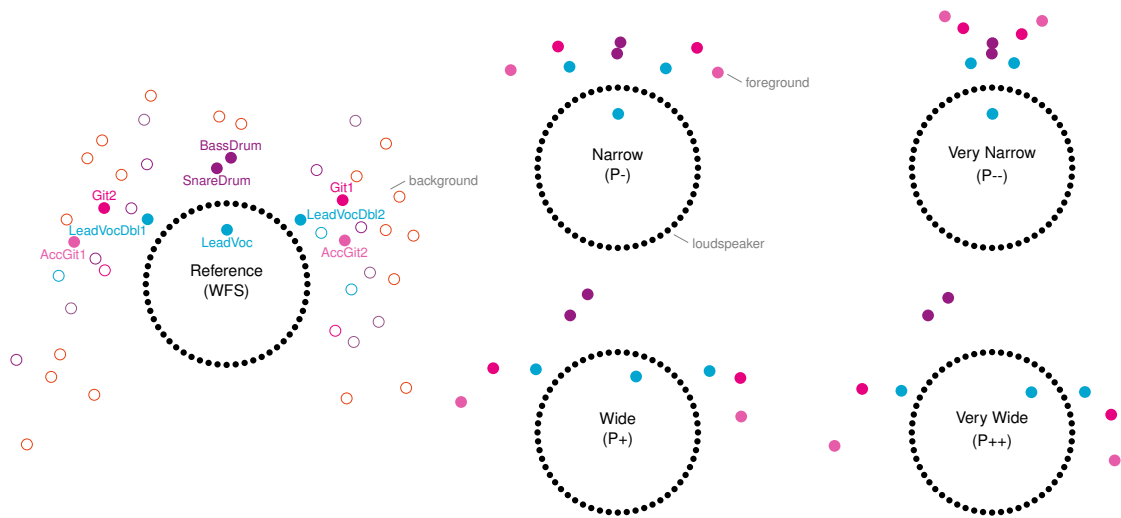


Figure 3.3: Arrangement of the five different object-based mixes for the WFS system. The positions of the different foreground elements (vocals, drums, guitars) are shown by the filled points relative to the circular loudspeaker array for the five different cases. For the reference mix the sound objects belonging to the background are indicated by the open circles, and omitted for the other conditions as they remained the same.

consists of a center foreground base a little narrower than stereo. In the wide (P+) mix, the foreground objects are gently pulled apart from the center of the scene, retaining an appropriate and symmetrical impression. Hence, the lead vocal and drum tracks are shifted inversely, with the drums on the left and lead vocals on the right side. In the very wide (P++) mix, some guitar parts finally appear from behind the listeners and the lead vocal from the right of the listener. The background part of the piece, including reverbs and delays, remain at their reference position in every WFS mix.

Compression There were alterations in both directions, more and less compression. An indicator for the amount of compression applied per instance is its “gain reduction”. Again starting from the reference mix (WFS), the variant containing more compression (C+) applies around twice as much average gain reduction per compressor. Therefore, the ratio of each compressor was doubled. In case of fixed-ratio compressors, or if the amount of compression was not sufficient yet, the corresponding threshold was lowered for the desired amount. The procedure was vice-versa for dialing in half the gain reduction for variant less compression (C-). Bypassing all compressor instances produced no compression (C--).

Reverb More reverb (R+) was introduced by increasing all the associated effect return levels by 6 dB. For less reverb (R-), those return channels were lowered by 6 dB accordingly. No artificial reverb (R--) corresponds to those effect returns muted.

The mixes are available in an object-based format as metadata Hold and Wierstorf (2016a) and signal feeds Hold and Wierstorf (2016b). The finished mix is available under Hold *et al.* (2016a). Note, that the finished mix starts at 84s of the original recording. In the listening test the first 30s of those mixes were used. The extract included pre-chorus and chorus, starting at a point consistent with musical phrasing with one bar leading into the pre-chorus. It was chosen as the chorus contains all the foreground instruments and as it contains a transition into the chorus, allowing participants to hear the processing characteristics in two slightly different settings.

Rendering and binaural synthesis

An open source WFS renderer (SoundScape Renderer Geier *et al.* (2008), with compensated distance dependent amplitude decay Hold *et al.* (2016b)) computed the loudspeaker driving signals that were then stored as sound files.

For the dynamic binaural synthesis Horbach *et al.* (1999) one binaural room scanning (BRS) file was created for every loudspeaker Wierstorf (2016) with a resolution of 1° utilizing high resolution head-related impulse responses Wierstorf *et al.* (2011). During playback the binaural synthesis software (SoundScape Renderer Geier *et al.* (2008)) convolved every BRS file with the corresponding loudspeaker driving signals, which were summed and returned as headphone signals. The binaural renderer updated the ear signals depending on the head orientation of the listeners.

The same loudness of the binaural signals for the different conditions was ensured by correcting the signals for a head orientation of 0° applying a loudness model (non-stationary Zwicker function of the Genesis loudness toolbox 1.2). For experiment 1 and experiment 2, the KEMAR dummy head was used to record the ear signals at the listening position for applying the loudness model.

Participants

21 participants (14 females; age range: 19–49; mean age: 32) were recruited for experiment 1. 42 different participants (29 females; age range: 19–73; mean age: 34) for experiment 2 and again 41 different participants (22 females; age range: 20–67; mean age: 29) for experiment 3. They self-reported no hearing loss or hearing disturbances. Informed written consent was obtained from each participant, and they received a financial compensation. The

study received ethical approval from the Technische Universität Berlin Ethics Committee (RA_01_20140422).

Procedure

Pairs belonging to one mix parameter or song were always grouped together, but the appearance of those groups was randomized. In each trial, participants were presented with a pairwise comparison of two temporally aligned clips of music, between which they could switch back and forth. For experiment 2 and experiment 3 playback stopped after the end of the 30 s long extract and an answer had to be given to advance to the next trial following an inter-trial interval of one second. In experiment 1, the whole songs were looped and participants allowed to listen as long as they wanted. They could always submit their answer before the end of the trial, given that they had heard a minimum of five seconds and had heard each of the two stimuli at least two times. The limiting of the presentation time to 30 s in experiment 2 and experiment 3 was motivated by the fact that more pairs were presented and to limit the position within the song the participants used for their judgements. Before the start of the experiment, participants practiced the paradigm twice with the experimenter. Here, another extract from the song was played and one of the tracks was presented at -6 dB.

In experiment 2 and experiment 3 participants completed a verbal survey asking for average daily hours spent listening to music and favorite music genres at the end of the listening test. Furthermore, participants were asked:

- 1) *When comparing a pair of stimuli, what did you pay attention to or which attributes of the mix triggered your decision?*
- 2) *Try to explain reverb, compression and equalization with respect to music production. Do you have expertise in sound mixing?*

These survey responses were recorded by the experimenter.

Statistical analysis

Suppose there is a number of musical pieces A, B, and C which should be assessed by listeners regarding their preferences. The advantage of the paired comparisons method to achieve this lies in its very few assumptions about the underlying process leading to the choices of the listeners. It is able to measure choices by the listeners, like circular triads where A is preferred over B, B over C and C over A. This can be a completely reasonable choice for stimuli that vary in different aspects. If instead a ranking of the stimuli or a preference rating on a scale is applied, it is already assumed that the rankings

lie on a one dimensional perceptual scale Kendall and Smith (1947). The pair-wise comparison circumvents this restriction and allows an analysis of the underlying dimensions afterwards.

An indication of a higher dimensional perceptual space is the systematic appearance of a high number of triads. Triads can also stem from inconsistent individual choice behavior and can occur in a non-systematic way when there is no agreement among the listeners. Counting the triads only provides a descriptive measure of the underlying choice process. In order to classify whether the appearance of triads is systematic, a statistical test is required. This can be achieved by fitting a Bradley-Terry-Luce (BTL) model Bradley and Terry (1952) to the data. A χ^2 goodness of fit test compares the estimated BTL model against an ideal saturated model for the paired comparisons. The BTL model holds and is not rejected, as long as the corresponding p -value does not drop below 0.1 Wickelmaier and Schmid (2004). If the BTL model holds it indicates that no systematic deviations from a one-dimensional perceptual space occur and estimates the choices of the listener on a ratio scale for that dimension Choisel and Wickelmaier (2007).

We used the BTL implementation in R (eba-package) after Wickelmaier and Schmid (2004). The BTL values are normalized to sum up to unity and present the probability that a given condition was preferred.

3.1.2 Results

For the evaluation, the ratings were aggregated over all participants in all three experiments. In order to test if a ratio scale could be obtained from the preferences, a BTL model was applied to the paired comparison data and then verified. In the following, every discussed BTL model fits the data, this means its p -value of the corresponding χ^2 -test never drops below 0.1 and the model is therefore not rejected. Only for the ratings on the changes in reverb in experiment 3 using binaural simulation the model did not hold. For comparison, the fitted data are nonetheless presented here.

In experiment 1, listeners rated their preferences for four different songs presented by stereo, surround and WFS. Figure 3.4 shows the estimated preference scores. They are normalized to 1 and indicate the ability of a given condition to win in a paired comparison to all of the other conditions. The results have a relatively large confidence interval, but show that there is a general trend across all songs that in general reproduction systems with a higher number of loudspeakers seems to be preferred. Especially stereo is always rated to be the least preferred reproduction system. This trend is further highlighted by the left graph in Fig. 3.6.

In experiment 2 and experiment 3 only the song Lighthouse was presented to the listeners, but the mixing parameters for the WFS condition were altered. The upper part of Fig. 3.5

3 Listening tests on sound quality

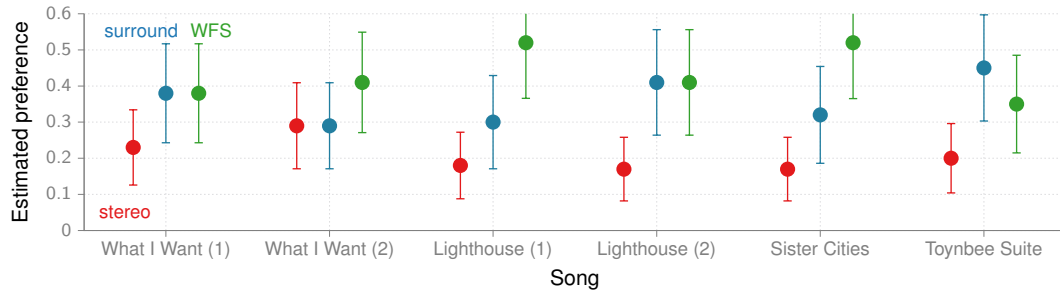


Figure 3.4: Average listening preference for stereo, surround, and WFS for different songs of popular music. The ratio-scale preference is estimated from paired-wise comparisons and shown together with 95% confidence intervals.

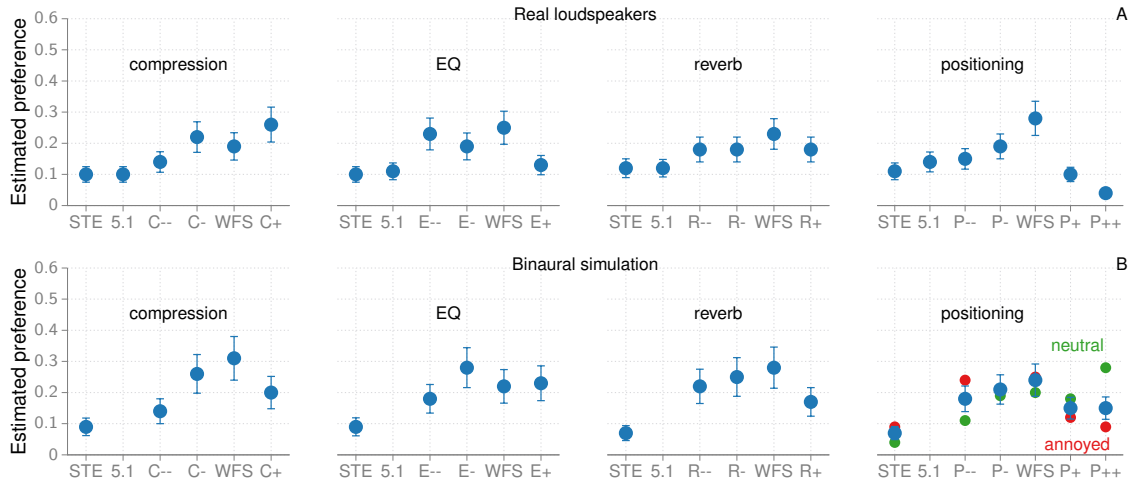


Figure 3.5: Average listening preference for stereo (STE), surround (5.1), and WFS with different mixes of a popular music song. The mixes for WFS changed in the denoted dimensions of *compression*, *EQ*, *reverb*, and *positioning*. Each processing was applied more (+), less (-) and bypassed (--). The green and red points for *positioning* represent the results for the two identified listener groups in the case of binaural synthesis. The ratio-scale preference is estimated from paired-wise comparisons and shown together with 95% confidence intervals.

shows the results from experiment 2 performed with loudspeakers in a room. The results are presented for the four different mixing parameters that were changed. It can be seen that in most cases the reference WFS mix (WFS) was the most preferred stimulus, only in the case of EQing and compression it was at a similar level or less preferred than the version for which the corresponding mixing parameter was more or less pronounced. For compression, EQ, and reverb, stereo and surround were equally the less preferred reproduction systems. Only in the case of the mixing parameter positioning, the wider conditions were less preferred.

Experiment 3 repeated experiment 2, but this time applied binaural simulation of the loudspeakers and omitted surround as a condition. The results are summarized in the lower graph in Fig. 3.5.

In order to assess the influence of dynamic binaural simulation instead of actual WFS reproduction, we evaluated the difference between the ratings happened with each reproduction method. The Wilcoxon signed rank test (implemented in R) compares the observations pairwise and its p -value (H_0 : difference between pairs of observations is zero) indicates significantly different ratings. We found significantly different ratings for positioning ($p < 0.01$). The test indicates no significant differences for the ratings of EQ ($p = 0.28$), compression ($p = 0.28$) and reverb ($p = 1$). Comparing all results in total indicates a difference between WFS and its binaural resynthesis ($p = 0.03$).

Further analysis highlighted that for experiment 3 there was a systematic disagreement between participants for the ratings of the positioning conditions. The disagreement was observed by calculating Kendall’s coefficient of concordance W (Legendre, 2005). It ranges from 1 (maximum agreement) to a minimum very close to zero, which is $W_{\min} = -0.02$ in our scenario. The corresponding p -values indicate how likely the agreement between judges is by chance, derived from a χ^2 -test. For all observations regarding positioning, a relatively low agreement of $W = 0.07$ ($p < 0.01$) was found, which indicates that there might be a systematic disagreement between the participants.

To analyze this disagreement, the participants were split into different groups based on the survey. For Question 1, 21 participants reported they were dissatisfied when lead tracks—especially the lead vocals—were shifted outside the center. The remaining 20 participants did not report any attributes directly related to positioning. This leads to grouping *dissatisfied* and *neutral*.

Regarding group effects, the likelihood ratio of individually calculated BTL models reveal whether two groups of subjects rated differently. This compares whether the combination of both group model likelihoods is significantly higher than the likelihood of the model calculated from the entire population, as this distance is approximately χ^2 -distributed.

For the two groups of participants that reported to be either neutral or dissatisfied with laterally shifted lead vocals, a significant difference was found ($p < 0.01$, from the corresponding χ^2 -distribution). This difference was quantified via Kendall’s rank correlation coefficient τ which describes the ordinal association between paired group outcomes. Correlation between participants neutral or dissatisfied with laterally shifted lead vocals is $\tau = 0.15$.

Both, experiment 1 and experiment 2 allow for a comparison of stereo, surround, and WFS. In experiment 1 this was achieved by averaging over the trials for the six different

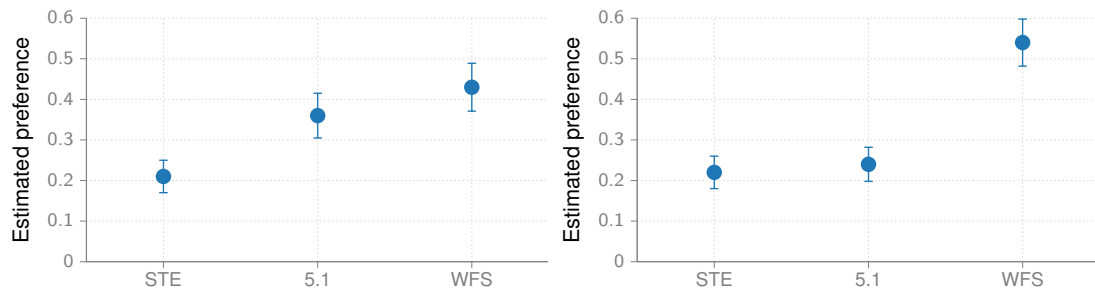


Figure 3.6: Average listening preference for stereo (STE), surround (5.1), and WFS in experiment 1 (left) and experiment 2 (right) The ratio-scale preference is estimated from paired-wise comparisons and shown together with 95% confidence intervals.

song excerpts and shown by the left graph in Fig. 3.6. In experiment 2 it was achieved by averaging over the trials for the four different mixing parameter changes, ignoring all the comparisons to the non-reference WFS conditions, which is shown in the right graph of Fig. 3.6. It can be seen that the surround conditions are less preferred in experiment 2 than in experiment 1.

3.1.3 Discussion

Altogether the results show that listeners prefer audio reproduction systems that use more loudspeakers and allow for a greater envelopment for listening to popular music. The differences between the systems seem to be more important than changes to the actual mixes. This is highlighted by the fact that only two of the 14 available WFS mixes were less preferred than stereo, even so the changes to the mixes along a specific parameter were relatively drastic. On the other hand those relatively drastic changes might explain why most listeners agreed with the mixing engineer in what is the preferred mix for WFS-


The differences in the ratings for surround between experiment 1 and experiment 2 are surprising and were not anticipated. The only difference between both experiments was the different context in which the paired comparisons took place. In experiment 2 a lot more WFS conditions and changes to the underlying mixes were presented to the listener compared to experiment 1. Why this has such a large effect is not clear yet.

For the modelling of the data it is important to get the binaural signals for the different conditions. Experiment 3 investigated if listeners would rate such binaural simulations in the same way they rate the actual reproduction systems using real loudspeakers. The outcome was more or less in agreement, only for the mixing parameter of positioning we found a significant difference. Here, the binaural simulation also led to two different groups

of listeners regarding the preferred mix of the arrangement. We think the positioning conditions might be especially critical as binaural simulations have problems with externalizations. As most of the current binaural models are not handling externalization as a feature, we think that it should be fine to use the binaural simulations of the systems to model the actual listening test results.

The modelling is presented in Chap. 4.

3.2 Finding the sweet spot in 5.0 surround

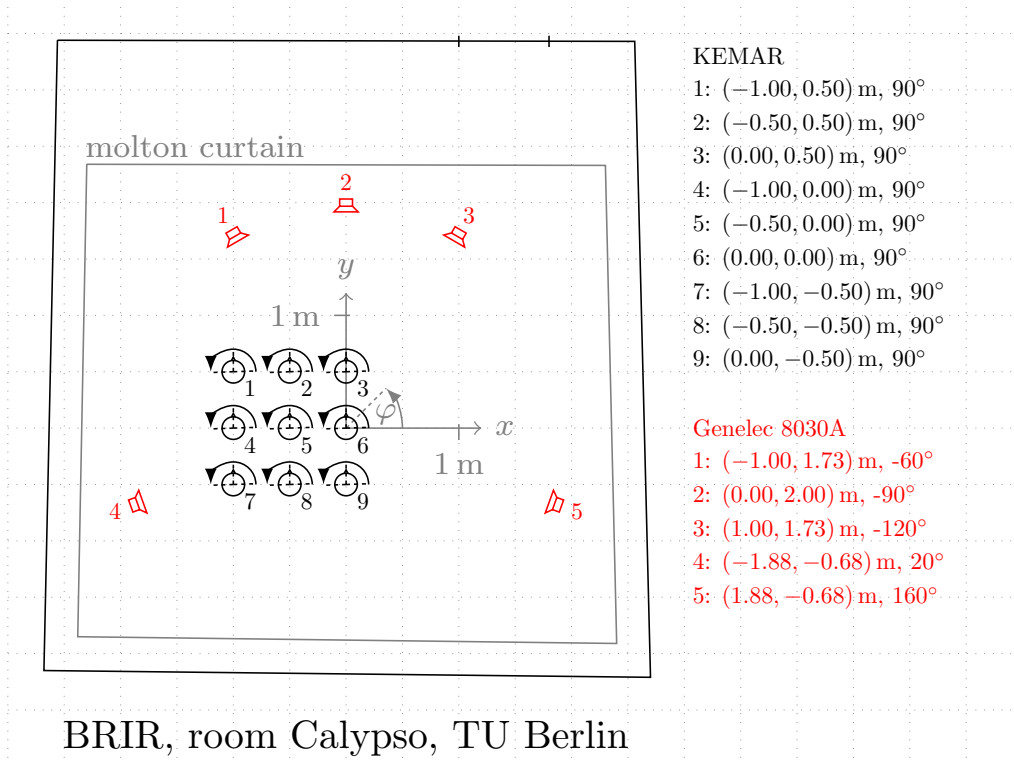
 **Published in database:** see online documentation

For audio reproduction systems the term sweet-spot is often used to describe a position or area of positions inside the listening area, in which the perception is as best as intended. For two-channel stereophony the sweet-spot area is in the center between the loudspeakers and is getting larger at positions farer away from the loudspeakers. For 5.0 surround the sweet-spot can depend not only on the reproduction, but also on the used recording technique. We tried to investigate the position and size of the sweet-spot for different recording techniques in 5.0, by letting the listeners decide which position they would like to sit to listen to the presented music. As their visual impression could lead to a bias of preferring for example center positions we conducted the experiment with and without visual orientation in the reproduction system.

3.2.1 Method

The experiment was performed in the studio like room Calypso in the Telefunken building of the TU Berlin. The experiment employed dynamic binaural synthesis in order to allow instantaneous switching between positions during listening. To accomplish this, BRIRs were recorded beforehand at nine different positions, see the description in D 1.3 and <https://doi.org/10.5281/zenodo.160761>. 26 test participants rated their preferred listening position out of the 9 positions for every recording technique. They did this first without visual feedback using a GUI (Graphical User Interface) that only showed nine buttons to which the stimuli were randomly assigned. In a second run they had a GUI showing a actual sketch of the listening setup, similar to the sketch shown in Fig. 3.7.

As music material seven different simultaneous recordings of the piece “Maurerische Trauermusik K.477” of W. A. Mozart were used. The recordings differed in the applied recording technique, which are listed in Tab. 3.1 and were done at the ORF (Austrian Broadcast, Vienna) and the piece was played by the Radio Symphony Orchestra Vienna. For more



KEMAR

- 1: $(-1.00, 0.50)$ m, 90°
- 2: $(-0.50, 0.50)$ m, 90°
- 3: $(0.00, 0.50)$ m, 90°
- 4: $(-1.00, 0.00)$ m, 90°
- 5: $(-0.50, 0.00)$ m, 90°
- 6: $(0.00, 0.00)$ m, 90°
- 7: $(-1.00, -0.50)$ m, 90°
- 8: $(-0.50, -0.50)$ m, 90°
- 9: $(0.00, -0.50)$ m, 90°

Genelec 8030A

- 1: $(-1.00, 1.73)$ m, -60°
- 2: $(0.00, 2.00)$ m, -90°
- 3: $(1.00, 1.73)$ m, -120°
- 4: $(-1.88, -0.68)$ m, 20°
- 5: $(1.88, -0.68)$ m, 160°

Figure 3.7: Setup of the BRIR recordings done in room Calypso. The nine measurement positions shown here are used in this listening experiment.

details on the applied microphones and for downloading all recordings have a look at Wittek (2015).

Table 3.1: Different recording techniques used and their corresponding abbreviations.

| Abbreviation | Recording technique |
|--------------|---------------------------------|
| Rec. 80 | Stereo + C + NHK |
| Rec. 81 | Decca-Tree + NHK |
| Rec. 82 | OCT + NHK |
| Rec. 83 | INA5 (Brauner ASM5) |
| Rec. 84 | Schoeps KFM 360 + DSP-4 KFM 360 |
| Rec. 85 | OCT-Surround |
| Rec. 86 | Soundfield MKV + SP 451 |

3.2.2 Results

Figure 3.8 summarizes the results across all listeners and recording techniques, showing only differences between the conditions with and without visual feedback. The results for the different recording techniques are summarized in Fig. 3.9. For more details on the experiment have a look at Schultze (2016).

3.2.3 Discussion

It can be seen in Fig. 3.8 that there seems to be a different preference for the preferred listening position depending on if the listener knows where it is actually positioned or not. In addition, it can be seen that there is no highly preferred single position. Both with and without visual feedback the four positions 2, 3, 5, 6 are preferred with a stronger preference for the center for the case of no visual feedback. The biggest difference between both presentations is observable at position 5 which was preferred more than three times as much for the case of visual feedback. It is not easy to see why this is the case. It might just reflect that the listeners thought that a listening position in the center of the available options was a good idea.

The biggest difference between the single recording techniques is that for some (Rec. 81 and Rec. 85) only positions near the center loudspeaker are selected at all, whereas for other like Rec. 84 or Rec. 86 the selected positions are more equally spread in the whole listening area.

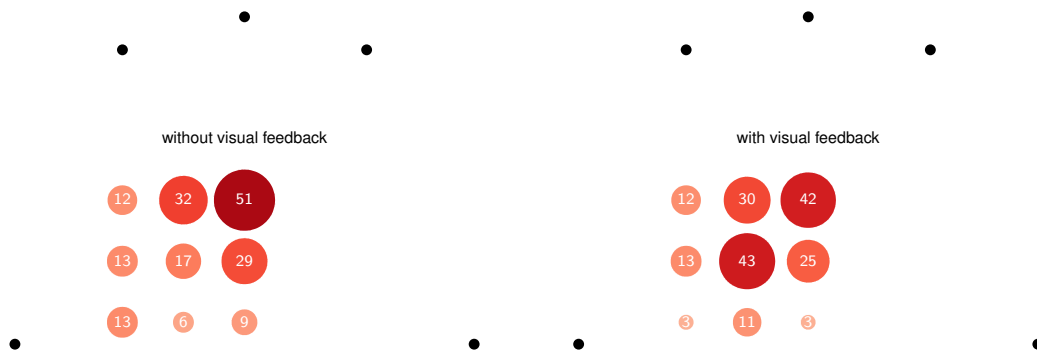


Figure 3.8: Number of chosen preferred positions summed over all listeners and recording techniques. On the left side the results without visual feedback and on the right with visual feedback about the actual listening position are shown.

3.3 Combined PC and MDS approach for listening preference

In a first formal test implementing the combined PC- and MDS-approach, a number of spatial audio conditions have been presented to listeners, playing out individual sound sources via different sound reproduction systems. Here, conditions similar to the ones used in the colouration assessment described in D6.2.2 have been used, so as to connect these tests with the already available data. Applying a single graphical user interface (GUI) for delivering votes of preference and perceptual difference between differently presented musical pieces, both PC-type data for analysis with a Bradley-Terry-Luce model and difference-scaled data for MDS were obtained.

3.3.1 Methods

20 listeners (aged between 20 and 30) took part in the experiment. Their hearing ability was verified using pure tone audiometry, and 18 of them were found to have normal hearing.

The same stimuli as in a previous WFS colouration experiment were used. See the online documentation on the previous test. All of them were binaural simulations of a circular loudspeaker array with a diameter of 3 m and a varying inter-loudspeaker distance between 67 cm and 1 cm. In addition, a single loudspeaker placed at the position, where the virtual point source was placed in WFS and the same loudspeaker with a high-pass filtered version of the audio material were included. As audio material, speech or music was presented as a single-channel signal processed as a point source placed in front of the listener. In order to avoid changes in colouration resulting from the employed non-individual HRTFs, the binaural synthesis used for presentation was not dynamic (that is, used without head tracking).

In a first session, the listeners completed a paired comparison preference test for both speech and music. After a short break, they repeated the experiment, but this time they were asked to rate the perceived difference between the conditions A and B on a continuous scale using a slider.

The results of the paired comparison test were analyzed with the Bradley-Terry-Luce model as introduced in Sec. 3.1.1.

The difference scaling data was analyzed in terms of multi-dimensional scaling using MATLAB[®], and a STRESS-value after Kruskal was calculated in order to estimate the number of underlying dimensions.

3.3.2 Results

The lower graph of Fig. 3.10 shows the average preferences for the different conditions as estimated by the Bradley-Terry-Luce model. The top graph in Fig. 3.10 shows the ratings from an earlier MUSHRA-experiment, where listeners rated the perceived colouration. It can be seen that the ratings for colouration and preference correlate quite well. In addition, the only difference between music and speech seems to be a more pronounced perception of colouration for music, which is expected as the artifacts of WFS are more pronounced at high frequencies.

The number of dimensions was determined based on evaluating the STRESS according to Equation 3.1. There, the values d_{ij} correspond to the difference values obtained between stimuli i and j from the test, and \hat{d}_{ij} are the distances obtained from the estimated configuration for the chosen dimensionality.

$$STRESS = \sqrt{\frac{\sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2}} \quad (3.1)$$

From the plot shown in Figure 3.11, it can be observed that a number of three dimensions leads to STRESS values around 0.1. A “knee” point as often used to select the number of dimensions was observed for two dimensions. Due to the better agreement indicated by the lower STRESS value around 0.1, three dimensions were ultimately cho-

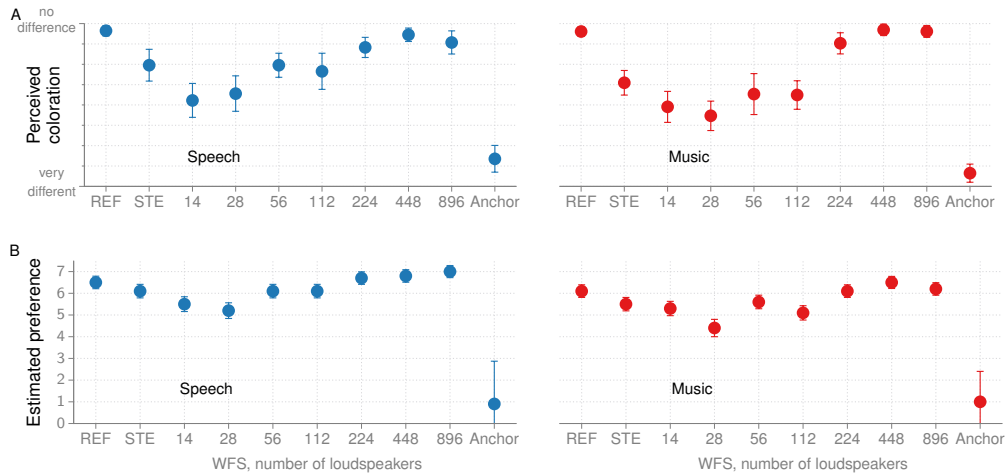


Figure 3.10: **A:** Average ratings on perceived colouration together with 95% confidence interval. **B:** Estimated preference values together with 95% confidence intervals. Both for varying number of available (WFS) channels, stereo (STE), a single virtual point source (REF) and its high-pass filtered version (Anchor).

3.3 Combined PC and MDS approach for listening preference

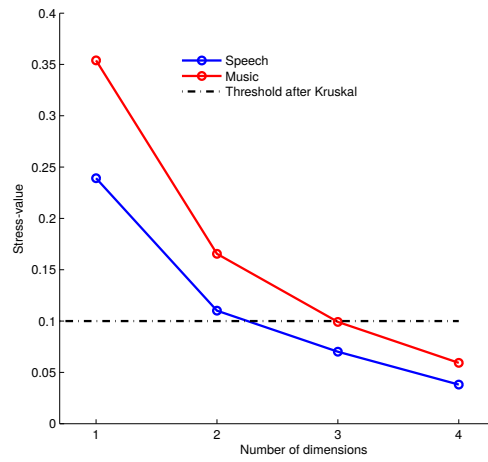


Figure 3.11: Selection of number of dimensions based on STRESS values.

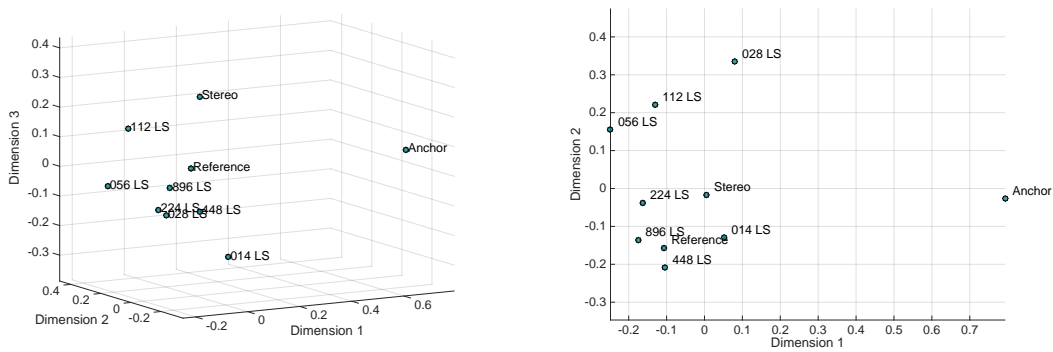


Figure 3.12: Multi-dimensional scaling for music as stimulus. Left: 3D view. Right: Dimensions 1 and 2.

3 Listening tests on sound quality

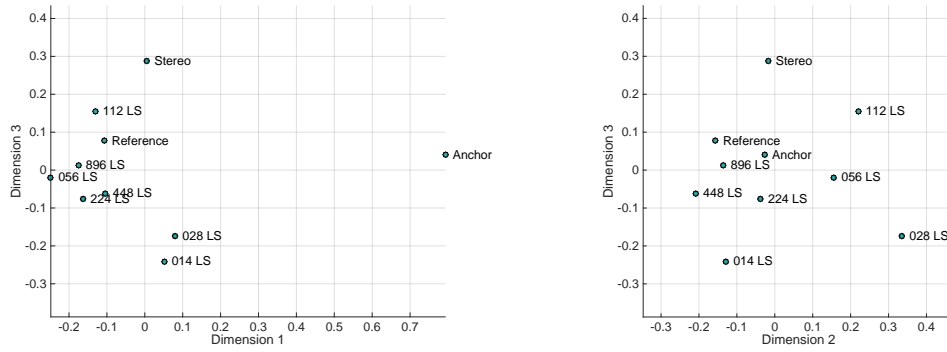


Figure 3.13: Multi-dimensional scaling for music as stimulus. Left: Dimensions 1 and 3. Right: Dimensions 2 and 3.

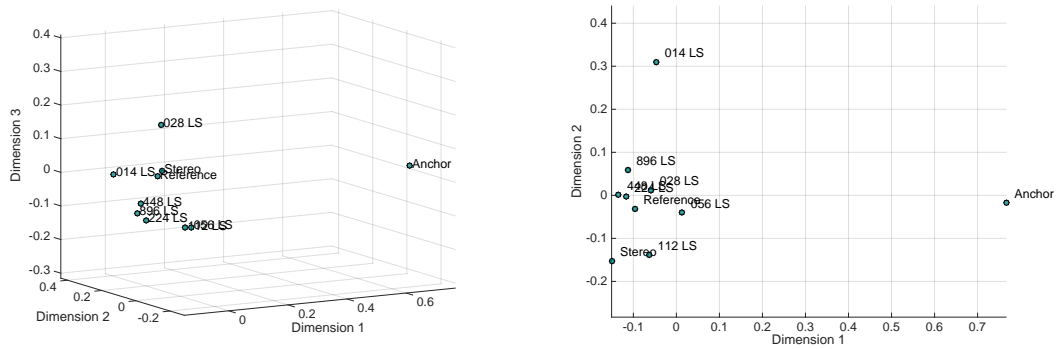


Figure 3.14: Multi-dimensional scaling for speech as stimulus. Left: 3D view. Right: Dimensions 1 and 2.

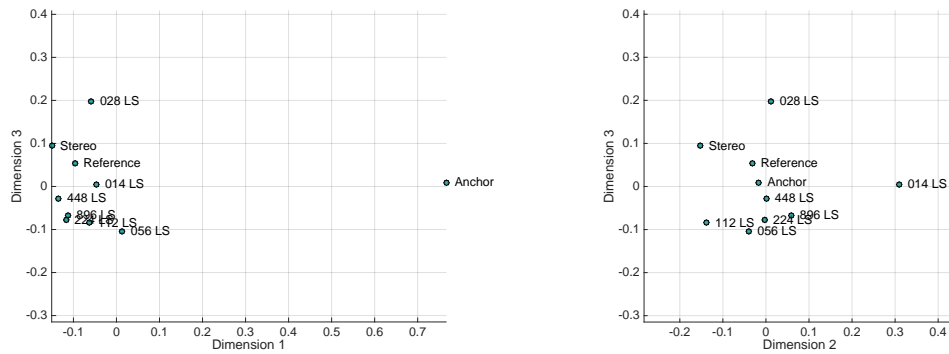


Figure 3.15: Multi-dimensional scaling for speech as stimulus. Left: Dimensions 1 and 3. Right: Dimensions 2 and 3.

sen.

The resulting representations of stimuli for music and speech in three dimensional configuration as well as for planes with two dimensions each are shown in Figures 3.12 to 3.13, and Figures 3.14 to 3.15, respectively.

3.3.3 Discussion

The resulting dimensions were difficult to interpret in terms of the underlying audition-related properties. This is in contrast to the work described for example by Brügger in Brügger (2001a,b), where a set of nameable dimensions for colouration could be determined. It is noted that in Brügger (2001a,b), attribute scaling was employed, while in the method selected for our case, MDS-data was used.

Due to the difficulties to interpret the data, it became clear that a direct elicitation of attributes would be needed to ensure the feasibility of later modelling. Here, the balance between the resources to be spent and the actual goal of considering a number of different features of TWO!EARS for modelling had to carefully be reflected.

For sensory evaluation, test subjects are needed that dispose of a certain level of expertise,

to reach the desired validity and reliability of the resulting dimensional data. This is typically achieved by dedicated pre-test training and prior recruitment of subjects, see for example Mattila and Zacharov (2001), Zacharov and Lorho (2006), Wältermann *et al.* (2010), Wierstorf *et al.* (2013). Considering our data and the limited selection and pre-training process used, it became clear that for a direct elicitation of attributes, the test subject panel was not “expert” enough so as to deliver the required agreement and thus resolution of the test results. Hence, it was concluded that more resources in the project would be needed for this work to address sensory evaluation for feature-identification. This result was anticipated in some way based on the literature, and one primary issue we faced was the fact that a solid listener panel was difficult to build up within the available time, also considering the split of working efforts between the different involved partners. As a consequence, it was decided to primarily focus the modelling activities on the dedicated creation of scene-based (that is, object-based) stimuli according to the parallel work stream I., also because of the very successful tests in this area when comparing different spatial audio systems.

3.4 Scene-specific quality assessment: Influence of source location and colouration combined

In D6.2.1, an experiment was described which investigated the perceived quality of a synthesised scene of three point sources (1 vocal and 2 guitars) that were processed towards various levels of colouration (Section 4.4, also reported in (Raake *et al.*, 2014)). The listeners’ grading for the various levels of degradation in colouration was collected via pairwise comparisons. One interesting finding was that the “unprocessed reference”, when not explicitly known to the listeners, was only ranked to be 8th out of the 10 stimuli. To further investigate this tendency in the non-reference quality judging situation, another experiment was designed and conducted, which extends the previous one in that another key aspect of spatial audio quality was additionally introduced - source location. It was expected that this extension would enable us to glimpse the consequences, if any, of combined manipulation of source colouration and location in this non-reference quality evaluation task. The long-term goal was to link the resulting preference data with the existing localization and colouration models for feature-based quality prediction.

3.4.1 Experimental design and procedure

The experiment was designed fundamentally in the same manner as the previous one, in that the same sources were used, and the same unrevealed “reference” scene configuration was used in pairwise comparisons - vocal at the centre, and two guitars from the left and right. Figure 3.16 shows this configuration.

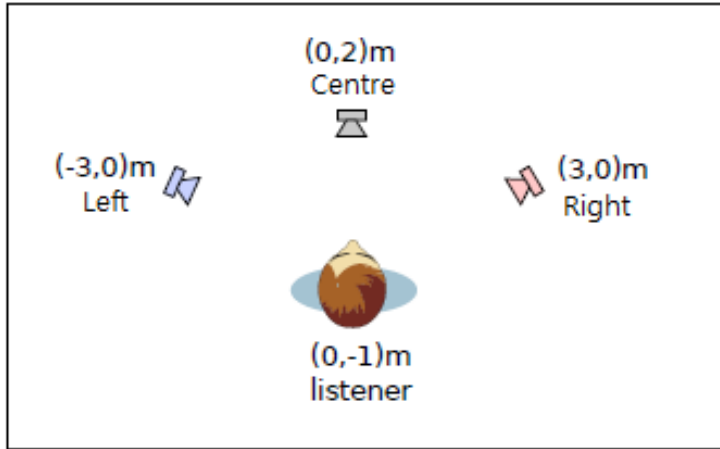


Figure 3.16: Synthesised source positions for the experiment, following (Raake *et al.*, 2014).

The number of stimuli for the pairwise comparison had to be limited at a maximum of 12, considering the resultant number of pairs to be evaluated and the required test duration. Therefore, it was decided to manipulate the source locations by switching their positions, rather than placing them at arbitrary positions. Also, not every possible combination of source positions could be considered, because of the limitation in the number of stimuli. Instead, some combinations were chosen to allow for as many variations as possible, whilst disregarding symmetric layouts (for instance, vocal at the centre - guitar 1 on the left - guitar 2 on the right was considered the same as vocal at the centre - guitar 2 on the right - guitar 1 on the left). The following table shows the synthesised positions of the three sources.

| Vocal | Guitar 1 | Guitar 2 | Abbreviation (Vocal-Guitar1-Guitar2) |
|--------|----------|----------|--------------------------------------|
| Centre | Centre | Centre | CCC |
| Centre | Left | Left | CLL |
| Centre | Right | Left | CRL |
| Left | Left | Right | LLR |
| Left | Right | Right | LRR |
| Right | Centre | Centre | RCC |

Colouration was applied only to the two guitar sounds. For the reason described above, not all levels of colouration used in the work of Raake *et al.* (2014) could be introduced. Colouration corresponding to “condition 3” in the previous experiment of Raake *et al.* (2014) was applied. This led to 2 levels of colouration per each of the location combinations in the table above. Therefore the total combinations of location and colouration can be described as follows (in the order of Vocal-Guitar 1-Guitar 2):

Ccc, CCC, Cll, CLL, Crl, CRL, Llr, LLR, Lrr, LRR, Rcc, RCC,

where the abbreviations are as described in the table, the capital letters denote unprocessed sources, and the small letters denote the “degraded” sources with colouration applied.

The binaural synthesis was done with the original sources used by Raake *et al.* (2014), and with the scripts to simulate WFS in various configurations with SFS Toolbox version 2.1.0. The 12 samples were used in the pairwise comparison experiment, leading to a total of 66 pairs to be compared. Each pair was presented twice. 12 listeners participated in the test. Figure 3.17 shows the user interface for the test.

The interface was designed using the Web Audio Evaluation Tool (Jillings *et al.*, 2016), which enables quick design of Python-based listening test interfaces using a web browser. Although it offers many convenient features such as randomisation and loudness control, it was later found out that this tool did not prevent accidental skipping of pages without indicating the answer. This led to invalid entities for some of the compared pairs (no win or loss). Therefore, for the analysis, only the first valid comparisons were selected (out of the two repetitions per pair).

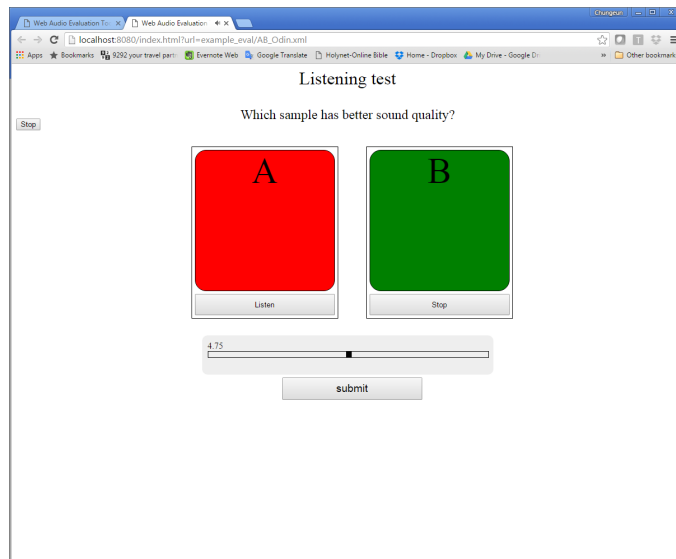


Figure 3.17: User interface for pairwise comparisons.

3.4.2 Results and analyses

A win-loss table was constructed from the comparison results. The Matlab function `fOptiPt.m` was used (Wickelmaier and Schmid, 2004) to estimate the preference rating with the Bradley-Terry-Luce (BTL) model. Figure 3.18 shows the preference ratings based on the pairwise comparison data. The stimulus configuration indicated on the horizontal axis is as described above.

The CRL configuration is rated the highest. This is the “unprocessed” setting comparable to the “reference” stimulus in the experiment of Raake *et al.* (2014).

Then from the win-loss table, Burstein’s (1988) formula was used to calculate the critical number of wins c' , required for a statistically significant difference between a pair.

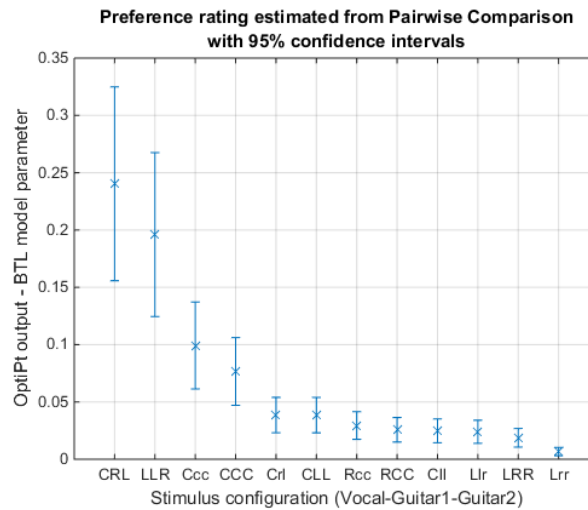


Figure 3.18: Preference rating converted from the pairwise comparison data through the BTL model, using `fOptiPt.m` function. See 3.4.1 for abbreviations.

| | | No. of losses | | | | | | | | | | | |
|-------------------|-----|---------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| | | Ccc | CCC | Cll | CLL | CrI | CRL | Llr | LLR | Lrr | LRR | Rcc | RCC |
| No. of wins | Ccc | 0 | 7 | 9 | 8 | 9 | 6 | 11 | 4 | 11 | 10 | 10 | 6 |
| | CCC | 5 | 0 | 8 | 7 | 9 | 4 | 9 | 3 | 12 | 8 | 10 | 9 |
| | Cll | 3 | 4 | 0 | 6 | 3 | 1 | 5 | 2 | 9 | 6 | 5 | 7 |
| | CLL | 4 | 5 | 6 | 0 | 5 | 3 | 6 | 2 | 11 | 10 | 6 | 6 |
| | CrI | 3 | 3 | 9 | 7 | 0 | 1 | 9 | 1 | 10 | 7 | 8 | 6 |
| | CRL | 6 | 8 | 11 | 9 | 11 | 0 | 11 | 10 | 12 | 11 | 10 | 12 |
| | Llr | 1 | 3 | 7 | 6 | 3 | 1 | 0 | 1 | 10 | 7 | 6 | 5 |
| | LLR | 8 | 9 | 10 | 10 | 11 | 2 | 11 | 0 | 12 | 12 | 11 | 11 |
| | Lrr | 1 | 0 | 3 | 1 | 2 | 0 | 2 | 0 | 0 | 2 | 2 | 6 |
| | LRR | 2 | 4 | 6 | 2 | 5 | 1 | 5 | 0 | 10 | 0 | 4 | 4 |
| | Rcc | 2 | 2 | 7 | 6 | 4 | 2 | 6 | 1 | 10 | 8 | 0 | 8 |
| | RCC | 6 | 3 | 5 | 6 | 6 | 0 | 7 | 1 | 6 | 8 | 4 | 0 |

The c' is calculated to be 9.358. This means that if a number of wins of a sample in a pair is equal to or larger than 10, then it can be claimed to be statistically significantly rated better than the other. The corresponding numbers on the win-loss table above are highlighted. It can be seen that the CRL and LLR samples were rated significantly better than the majority of the rest. Samples Cll, Lrr, and RCC were never rated significantly better in any of the comparisons.

3.4.3 Effect of colouration

If we group the stimuli into groups with the same location distribution of the sources, according to Fig. 3.18, it seems that the stimuli with colouration applied to the guitar sounds tend to be rated better than without colouration, except for two cases (CCC and RCC configurations). More specifically, the numbers of wins per pairs, in the order of how clearly the pair is distinguished, are as follows:

- CRL vs CrI – 11:1
- LLR vs Llr – 11:1
- LRR vs Lrr – 10:2
- RCC vs Rcc – 4:8
- CCC vs Ccc – 5:7
- CLL vs Cll – 6:6

The two most clearly distinguished pairs both have the two guitars separated to the left

and right, indicating that the colouration was more easily noticeable. Additionally, the comparisons for the later three pairs indicate that the applied colouration was perceived confusingly, sometimes with the colouration-applied stimulus more preferred. The difference between the first three pairs and the rest is that the auditory scenes were wider for the first three pairs. This implies that colouration might have been more difficult to distinguish when it is applied to sources that are placed closer to each other.

3.4.4 Effect of source location

If we separate the stimuli into two groups – with and without colouration applied, the estimated order of preference is

- CRL-LLR-CCC-CLL-RCC-LRR (without colouration)
- Ccc-Crl-Rcc-Cll-Llr-Lrr (with colouration)

The order of preference is not the same in the two groups, making it difficult to conclude a trend due to the source location configuration alone. The differences between the stimuli within a group, however, seem to be larger and more often significant when colouration is not applied. In other words, the location configurations were relatively less distinguishable in terms of preference when colouration was applied.

Amongst those configurations without colouration applied, a strong preference is seen for the two stimuli with the two guitars separated to the left and right (CRL and LLR). This cannot be generalised to the other group (with colouration). However, as already described, the differences between the stimuli with colouration are too small to extract any location-related tendency. Still, it is interesting to see that when colouration was applied, the Ccc stimulus was rather clearly preferred to the rest (in the group).

3.4.5 Discussion

One of the initial aims of this study was to investigate the effects of manipulating the source locations in addition to applying colouration, in comparison to the findings of Raake *et al.* (2014). However, it was found later that the colouration applied to the guitar stimuli in the current experiment, supposedly using the same binaural synthesis of WFS configuration as in the study of Raake *et al.* (2014), was in fact not exactly the same as what was applied to the stimuli in their experiment. More specifically, the guitar sounds with the colouration applied by means of the “newer” version of the SFS Toolbox had a clear lack of low-frequency components, whereas the guitar sounds with colouration applied in the experiment of Raake *et al.* (2014) were not so clearly distinguishable from those

without colouration. Therefore it was not possible to directly compare the results from this experiment to those from the study of Raake *et al.* (2014). However, the current experiment is valid on its own, with the following findings:

- The unrevealed reference was rated the most preferred, in contrast to the previous experiment of Raake *et al.* (2014), in which the reference was ranked 8 out of 10
- Colouration seems to be more easily distinguished when the sources were separated rather than closer together

Overall the results do not seem straightforward enough to be modelled with the TWO!EARS framework. This indicates the complexity of the preference-judging process, especially as another quality-affecting aspect of auditory scene is introduced. More subjective data in more various but controlled situations, for instance with quantifiable (if ever possible) degradation of attributes, would be required if a robust model is to be developed.

3.5 Concert hall recordings – preference and perceptual attributes

Conventional objective evaluation of room acoustics often involves prediction models that are derived from measurements of a number of physical properties of sound that can be matched to the listeners' perceptual grading of a certain attribute in a given listening space. Recent research works conducted by van Dorp Schuitman (2011) and van Dorp Schuitman *et al.* (2013), however, took a different approach in that attempts to predict some acoustical properties (reverberance, clarity, Apparent Source Width (ASW) and Listener Envelopment (LEV)) were made based on the internal representations of the binaural input signals. This fits well with the TWO!EARS framework, which motivated an investigation into the model's validity in other circumstances than what had been introduced in their own validation of the model. Also, in particular, the concert hall acoustics attributes were chosen as the evaluated properties because evaluation of concert hall acoustics can be considered as a non-reference situation, in which more often than not the listeners do not have any access or knowledge of an explicit and fixed "ideal listening experience". The aims of this investigation were accordingly set to the followings:

- To validate the above mentioned model (van Dorp Schuitman, 2011), referred to as the van Dorp Schuitman model hereafter) in terms of its ability to distinguish stimuli collected in a different environment
- To investigate how the prediction outputs of the model can be related to the preference of concert hall recordings in the non-reference evaluation

3.5.1 Experimental design and procedure

A set of binaural dummy head recordings of a piano sound, at different spots of an empty concert hall, was made available by the courtesy of Constant Hak (TU Eindhoven) and Remy Wenmaekers (Level Acoustics), only for the research purpose. The recordings were made at Muziekcentrum Eindhoven, a small concert hall with 490 seats. The playback of the piano was controlled via MIDI, so that exactly the same sound would play when the dummy head was moved to different spots for repeated capturing. A total of 6 recordings were made and used for the evaluation. Figure 3.19 shows the spots where the recordings were made: on the stage (K in the picture), R(row)02S(seat)06 (numbered 1 in the picture), R09S18 (2), R14S07 (3), R20S06 (4), and R24S06 (5).

For this experiment, a multi-stimulus test scheme was used instead of pairwise comparison (De Man and Reiss, 2013). In their work it was found that the multi-stimulus test revealed the same preference grading as the conventional pairwise comparison method, with a more detailed distinction in some cases. The use of this method enabled hiring of more listeners and asking for multiple attributes during the test session, thanks to the reduced time to conduct the listening tests, compared to the pairwise comparison. Figure 3.20 shows the user interface to collect the preference gradings. Here the listeners could listen to all the 6 stimuli by clicking the corresponding tabs on the scale, and were asked to move them according to their liking of each stimulus compared to the others. Once the “Preference” page is finished, the reverberance and clarity ratings were similarly collected on a separate screen. This enabled comparison of the model output to the subjective gradings, although

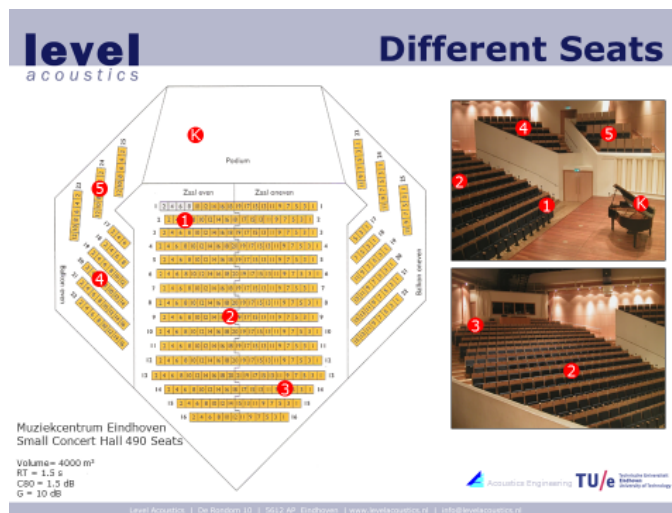


Figure 3.19: Diagram and pictures of the concert hall Muziekcentrum Eindhoven where the binaural recordings were made. Courtesy of Remy Wenmaekers (Level Acoustics).

ASW and LEV were not included due to time constraints. A total of 22 non-expert subjects took part in the test.

3.5.2 Results and analyses

The ratings were on a scale of 0 to 100. Because there was no explicit reference or anchor used in the experiment (in order to avoid any possible bias), the collected scores were “normalised” with respect to mean and standard deviation, according to ITU-R BS.1284-1.

The van Dorp Schuitman model, available as a black-boxed software package named RAA (Room Acoustic Analyzer), was provided by the courtesy of the author and run to get the model output data from the 6 stimuli. Four parameters were calculated – scaled reverberance (sRev), scaled clarity (sCla), scaled ASW, and scaled LEV (sLEV).

Firstly, Fig. 3.21 shows the subjective scores as box plots and the model outputs (asterisks) for reverberance. The model output values were multiplied by 100 to be on the same scale (of 0 to 100) as the subjective scores. The model output values were very close to each other, making it difficult to see the differences compared to the box plots. Therefore the model output was plotted separately on Fig. 3.22, without multiplication by 100. Although there is a difference in the resolution, it can be seen that the model output closely follows the tendency in the subjective scores.

Secondly, the subjective scores for clarity were compared to the model outputs in the same

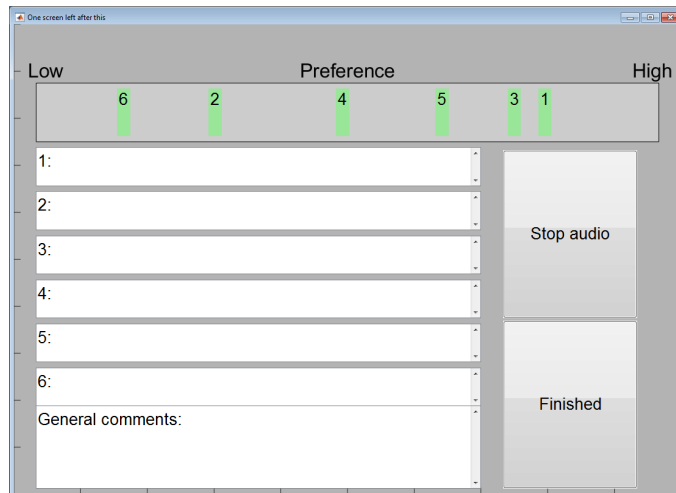


Figure 3.20: The user interface for the multi-stimulus test scheme. The listeners could listen to all the stimuli on the same page and graded their liking, or perceived attributes on a single scale.

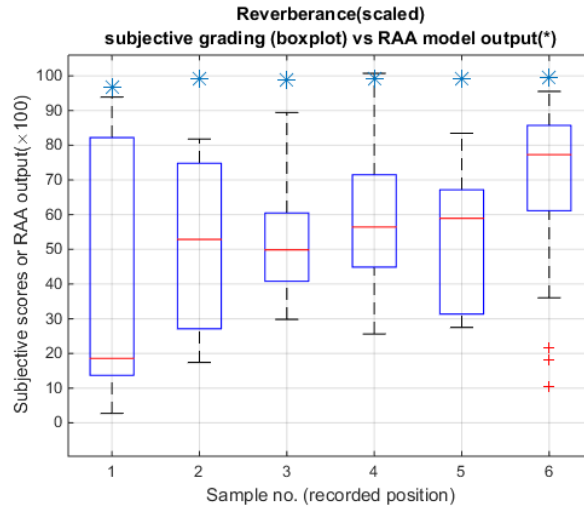


Figure 3.21: Subjective gradings of reverberance (box plots) and the RAA model outputs (asterisks, multiplied by 100 to be on the same scale).

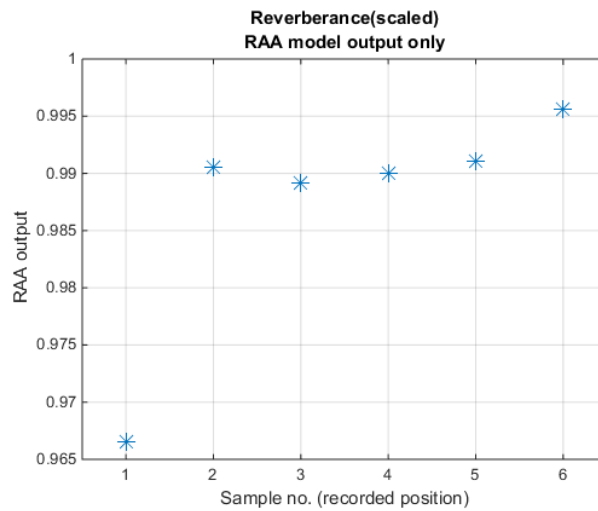


Figure 3.22: RAA model output of reverberance (sRev parameter in the software), without multiplication.

way in Fig. 3.23 and Fig. 3.24. Again the overall tendency seems to be predicted with the model output. Although the order of output scores is different for samples 2 to 4, it is expected that these differences are not statistically significant.

The two attributes above – reverberance and clarity – are the ones both with subjective scores and model outputs available. For the other two attributes ASW and LEV, subjective scores were not collected and only the model outputs are available. These are plotted in Fig. 3.25 and Fig. 3.26.

And for the overall preference, the box plot and the error bars in Fig. 3.27 and Fig. 3.28 show the subjective scores. Although more detailed statistical analysis is to follow, it seems that overall the recording on the stage was the most preferred, and the ones on the side balcony were the least preferred, as expected.

The preference rating was converted into a win-loss table shown below, as from pairwise comparison, for further analysis according to the suggestion by Burstein (1988).

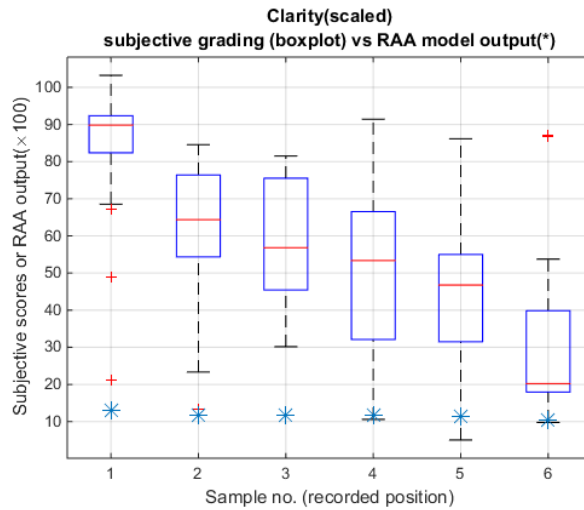


Figure 3.23: Subjective gradings of clarity (box plots) and the RAA model outputs (asterisks, multiplied by 100 to be on the same scale)

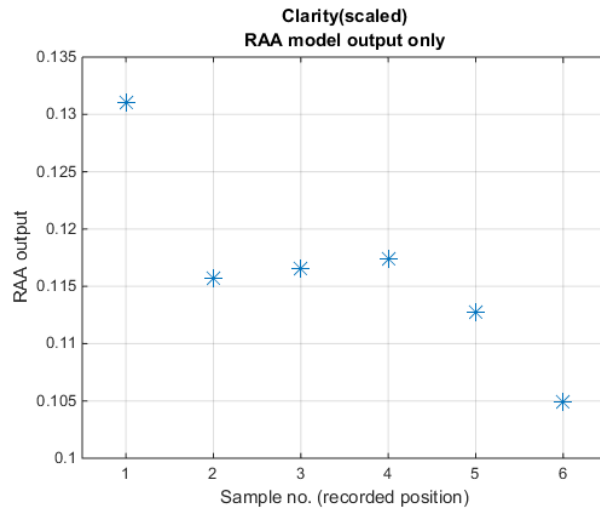


Figure 3.24: RAA model output of clarity (sCla parameter in the software), without multiplication.

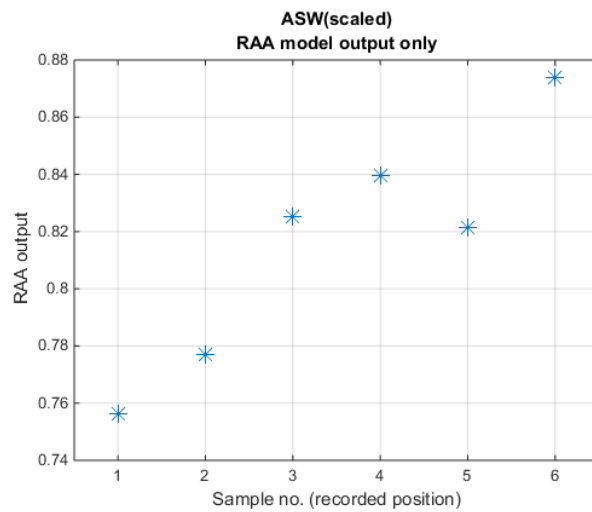


Figure 3.25: RAA model output of ASW (sASW parameter in the software).

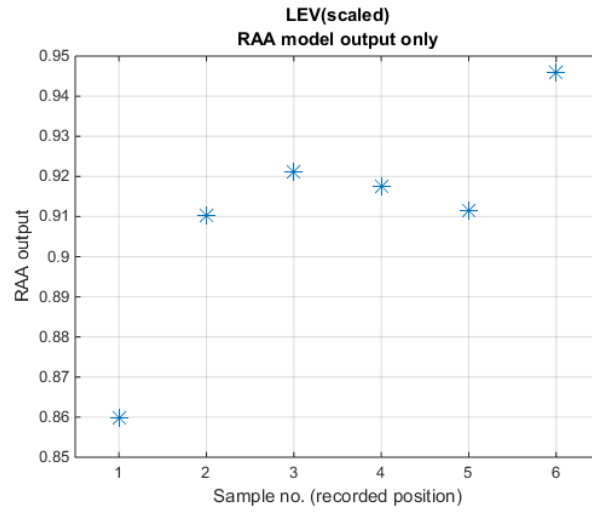


Figure 3.26: RAA model output of LEV (sLEV parameter in the software).

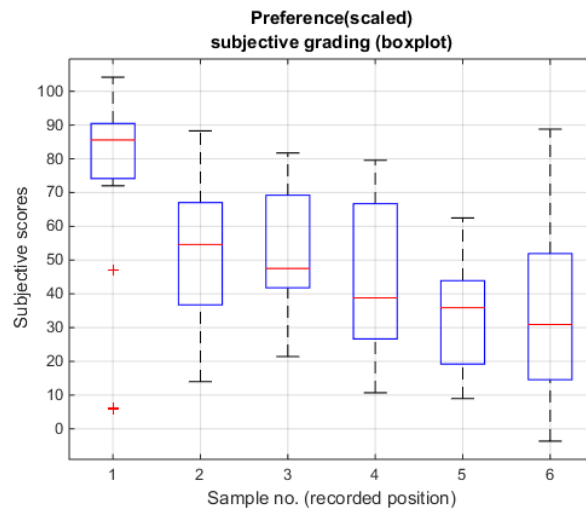


Figure 3.27: Box plots of preference grading.

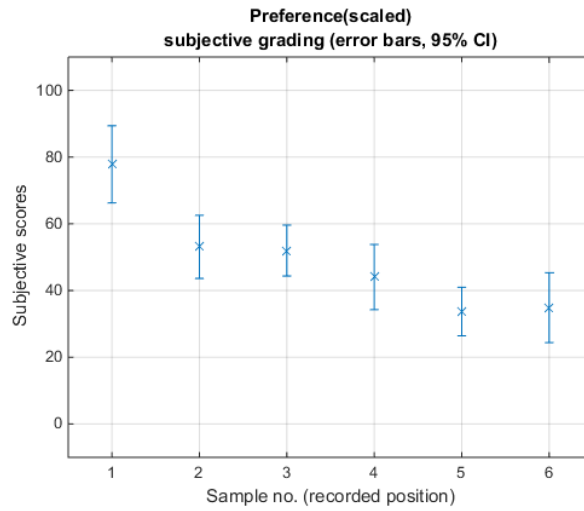


Figure 3.28: Error bars of preference grading.

| | | No. of losses | | | | | |
|--------|----------|---------------|----------|----------|----------|----------|----------|
| | | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | Sample 6 |
| | Sample 1 | 0 | 17 | 20 | 19 | 20 | 18 |
| | Sample 2 | 5 | 0 | 12 | 14 | 15 | 17 |
| No. of | Sample 3 | 2 | 10 | 0 | 15 | 17 | 17 |
| wins | Sample 4 | 3 | 8 | 7 | 0 | 17 | 14 |
| | Sample 5 | 2 | 7 | 5 | 5 | 0 | 11 |
| | Sample 6 | 4 | 5 | 5 | 8 | 11 | 0 |

The point of interest here is whether there is any significant difference in terms of preference ranking between a pair of samples. The null hypothesis then becomes H_0 : (per pair) the probability of a sample winning against the other is 0.5 (due to chance).

Assuming that the sample size is large enough for normal approximation, the number of wins of a sample against another, required for statistically significant difference, can be calculated from one of Burstein's equations to be 15.37.

This means that if a number of wins of a sample in a pair is equal to or larger than 16, then it can be claimed to be statistically significantly rated better than the other. The corresponding numbers on the win-loss table above are highlighted. It can therefore be said that sample 1 is perceived significantly better than all the others, sample 2 was significantly better than sample 6, sample 3 better than samples 5 and 6, and sample 4 better than sample 5. From this, it seems reasonable that the samples can be divided into 3 groups in terms of overall preference – (sample 1), (samples 2, 3, and 4), and (samples 5 and 6).

3.5.3 Discussion

The following conclusions can be made from the findings so far:

- The van Dorp Schuitman model was able to distinguish the different listening spots in terms of the four acoustical attributes
- The prediction of reverberance and clarity from the van Dorp Schuitman model correlated closely to the subjective gradings
- The stimuli could be divided into three groups in terms of the overall preference, with the one at the stage perceived to be the most preferred, and with the ones on the balcony to be the least preferred

The overall preference grading and groups reflect what one would expect in this listening situation: the stage recording was the most preferred and the balcony was the least preferred. Interestingly, comparison of the overall preference rating and the individual subjective or predicted outputs of the four parameters shows that the overall preference tendency is positively correlated only with the clarity. This implies that clarity contributed positively to the preference, as opposed to reverberance. Further research in this aspect would be desirable, because the source characteristics might have affected this tendency - the piano has distinctive attacks or onsets, for which clarity may be a good indicator, compared to bowing instruments such as violin, for example.

Although the remaining two attributes ASW and LEV still need further investigation as to whether the model outputs would match with the listeners' perception, it seems plausible at this stage to establish a framework where this model can be further validated against actual human perception and evaluation in various other situations. Since the van Dorp Schuitman model has many internal processing components in common with the TWO!EARS auditory front-end, it was decided to implement this model within the TWO!EARS framework, based on the published works. More details of the model implementation will be described under Section 4.

3.6 Compression of interaural level differences

The level-dependent nonlinearity of the basilar membrane operation implies that the binaural cues dependent on the signal level, such as the Interaural Level Difference (ILD), can be altered at the inner ear, depending on the presented ("reference") level. Especially, simulations using the DRNL filterbank implemented in the TWO!EARS auditory front-end showed some noticeable variations in the ILD and interaural coherence (IC) (Kim and Kohlrausch, 2016). In order to confirm whether the signal level actually affects the

perception of ILD as found in the simulation, a set of psychophysical tests were conducted to investigate into the strength of perceived ILD over various presentation levels. More specifically, the time-intensity tradeoff in sound lateralisation was investigated at various levels, as in a previous work by David *et al.* (1959), which was expected to indicate the relative strength of ILD. This would in turn help to validate the usefulness of the DRNL filterbank-based nonlinear AFE model to predict the level-dependent variations of some spatial cues.

3.6.1 Test method and procedure

A total of 9 paid subjects participated in the listening test. They were asked to listen to the stimuli presented over headphones and to indicate the laterality of the sound, on a screen of user interface designed with APEX software (Francart *et al.*, 2008). Figure 3.29 below shows a screenshot of the user interface.

The stimuli were generated based on a 2-kHz tone. Each stimulus had three short bursts of a 100 ms-long tone with a ramp of 5 ms, in 100-ms intervals. A total of 5 reference levels were introduced: 20, 35, 50, 65, and 80 dB SPL. The experiment session was divided into two parts, in which a stimulus with a fixed ITD was initially presented and the ILD was adaptively applied depending on the subjects' answers ("fixed ITD" case), and in which a fixed ILD was initially presented, and the ITD was adaptively applied ("fixed ILD" case).

For the fixed ITD part, the ITD values of $-600\ \mu\text{s}$ to $600\ \mu\text{s}$ in steps of $100\ \mu\text{s}$ were used for the initially presented stimuli. Depending on the answer of the subject indicating whether the sound is perceived to be on the left or the right side, the ILD was applied so that the perceived sound can be shifted towards the centre. The ILD was applied in decreasing

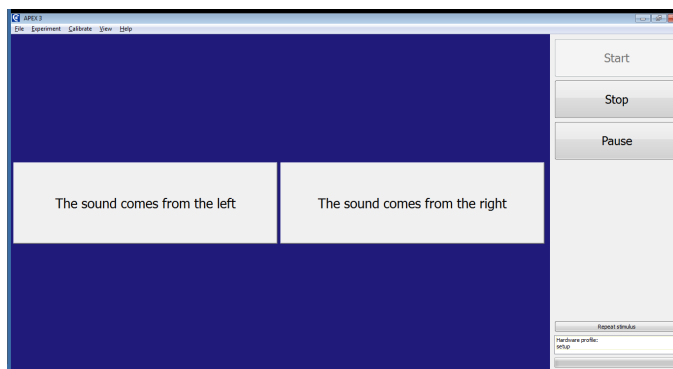


Figure 3.29: User interface for the adaptive psychophysical listening test to centralise the lateralised stimuli.

step sizes as the subject progressed with answers, with the minimum step size of 2 dB. For example, if the initial answer were “the sound comes from the left”, an ILD of +6 dB would be applied (+6 dB step), and if the next answer were “the sound comes from the right”, an ILD of +2 dB would be applied (−4 dB step). The point at which the perceived direction of the sound was shifted to the opposite direction was referred to as “reversal”. When two reversals happened, the average ILD was taken as the corresponding value to compensate the fixed ITD, and the next centralisation process started with another fixed ITD presentation. From some initial test runs, it was found that some subjects could not lateralise or centralise the stimuli even with an excessive ILD applied. Therefore, for the actual listening test, a so-called “ceiling” was introduced at ± 13 dB. When no reversal was made even after the largest amount of ILD was applied, at the next step the ILD was forcedly applied towards the opposite direction so that the stimulus was lateralised to the other side.

For the fixed ILD part, the initially presented ILD was from -12 dB to 12 dB in 2-dB steps. The ITD was adaptively applied depending on the subject’s answer, ranging from -1900 μ s to 1900 μ s. The step size started from 300 μ s, and after one reversal was reduced to 100 μ s. The ITD value for compensating the presented ILD was determined after three reversals. The ceiling was also introduced in this case, such that a reversal would always happen when the maximum amount of ITD (± 1900 μ s) was applied without any reversal.

3.6.2 Results and analyses

For each of the two parts (fixed ITD / fixed ILD), the ILD or ITD values were collected that were found to compensate the presented ITD or ILD. Firstly, the data points (ITD-ILD pairs) due to the reversals at the ceilings were identified and removed for analyses. The plots in Figs. 3.30 and 3.31 show the overall results from all the subjects in the fixed ITD / fixed ILD parts respectively, for each reference level. A 3rd-order polynomial fitting result is also shown for each plot, with the 1st-order coefficient (p3), the constant term (p4), and the goodness of fit information (R2). The variance of the ILD values in the fixed ITD part is large in general.

Then following the work of David *et al.* (1959), the results per individual subject were examined. From the scatter plot, the time-intensity tradeoff ratio was estimated per reference level as follows:

- A 3rd-order polynomial was fitted to the scatter plot
- The 1st-order coefficient was used as the ratio.

Figs. 3.32 and 3.33 show the results for one of the subjects, separately for the fixed ITD

3.6 Compression of interaural level differences

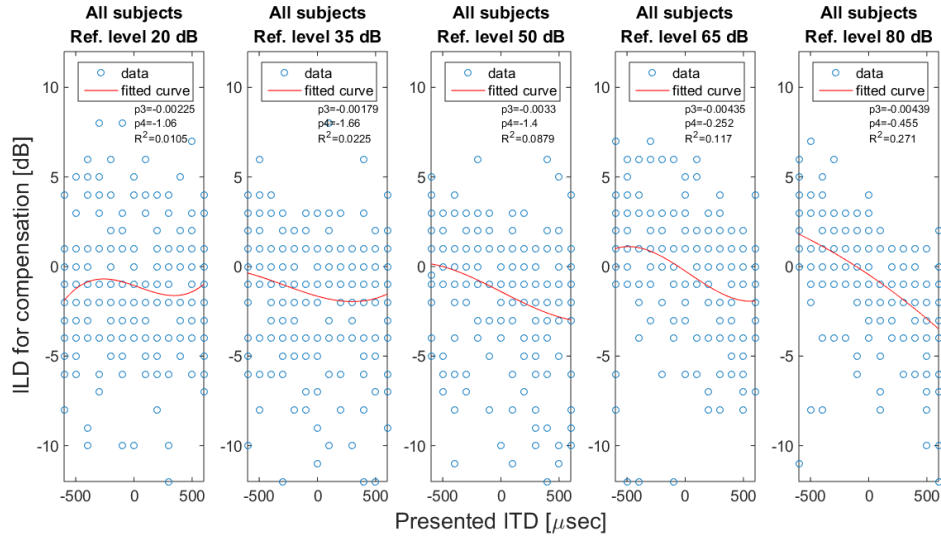


Figure 3.30: Data from all subjects, fixed ITD

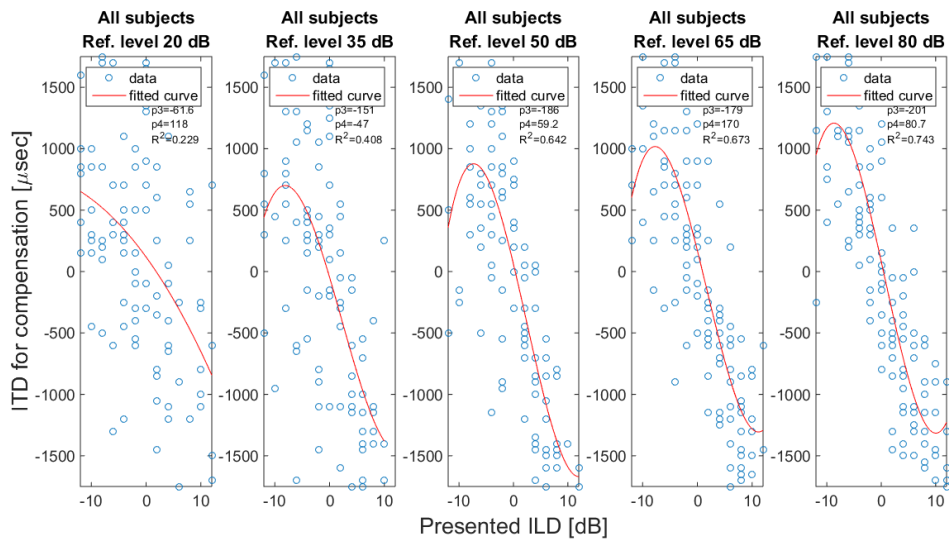


Figure 3.31: Data from all subjects, fixed ILD

and fixed ILD parts. The p_3 values were used as the estimated tradeoff ratio for further analysis across the subjects.

Figure 3.34 shows the 1st-order coefficients collected as described above, for the fixed ITD part. These values indicate the estimated amount of ILD in dB to compensate the presented ITD of $1 \mu\text{s}$. It seems to show a trend of increasing as the reference level increases. Figure 3.35 shows the box plots, along with the means and 95% confidence intervals, of the estimated ratios.

As the reference level increases from 20 dB to 35 dB, the mean and median decrease. Then they both increase as the reference level increases up to 65 dB, followed by a decrease at the highest tested reference level. This ratio (the coefficient) can be considered as the inverse of the strength of internally perceived ILD for the varying reference level: a decrease of the ratio means that the strength of ILD increases, and vice versa.

Figs. 3.36 and 3.37 show the 1st-order coefficients in the same way, now for the fixed ILD part. These values indicate the estimated amount of ITD to compensate the presented ILD of 1 dB.

The box plots and the error bars with means and 95% confidence intervals show that when the reference level increased from 20 dB to 35 dB, both the median and mean increased, and as the reference level increased further to 65 dB, they tend to decrease, except for the median for 35 dB to 50 dB reference level. As the reference level reached the highest tested value of 80 dB, the median and mean increased again. This ratio can be considered as the strength of the internal ILD – for example, increase of the ratio here means that

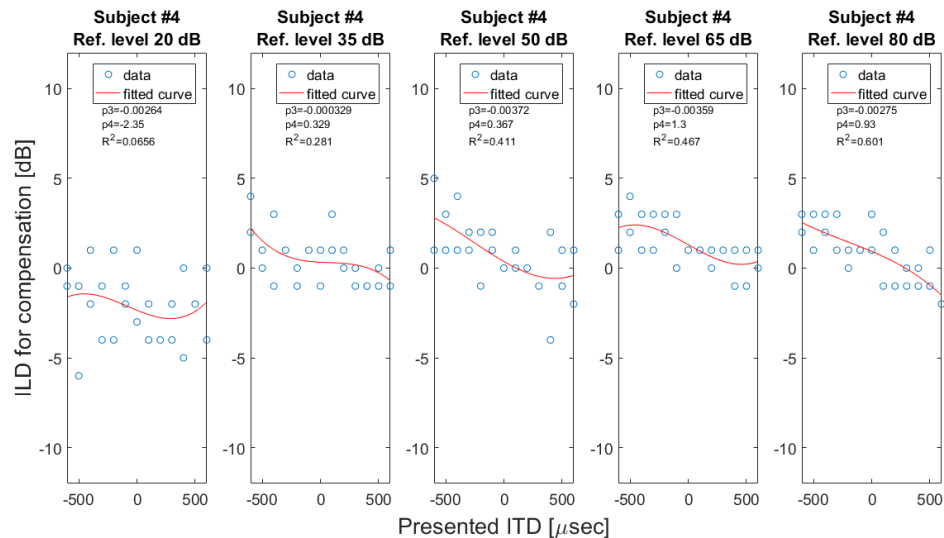


Figure 3.32: Data from one subjects, fixed ITD

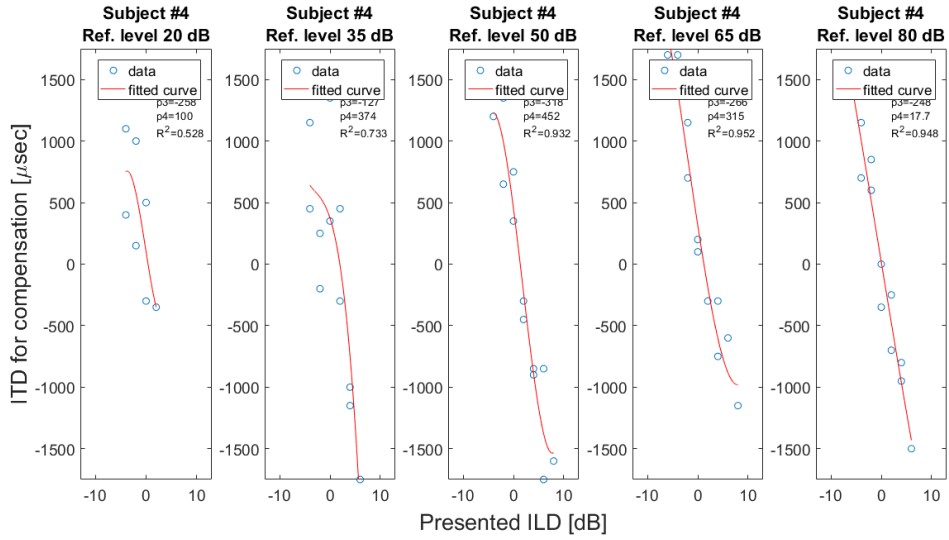


Figure 3.33: Data from one subjects, fixed ILD

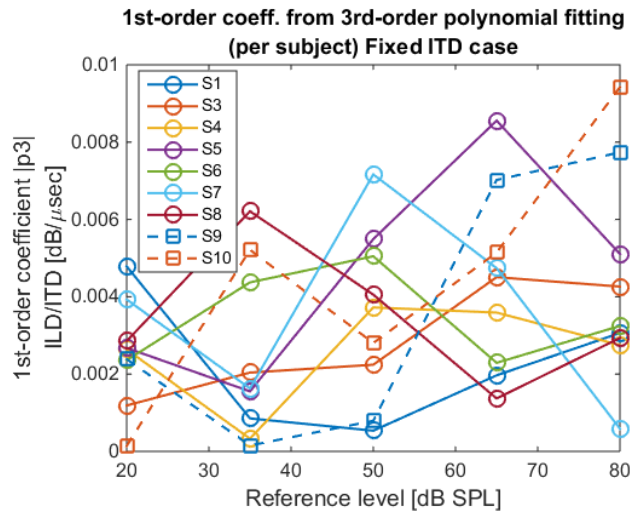


Figure 3.34: 1st-order coefficient from the 3rd-order polynomial curve fitting of individual dataset, per reference level, in the fixed ITD case.

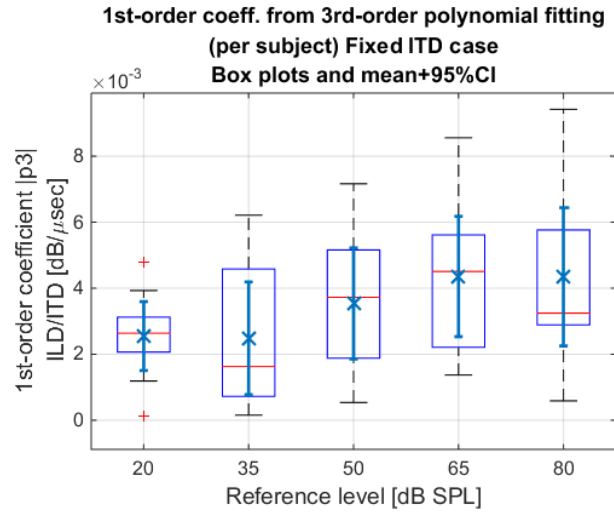


Figure 3.35: Box plots and mean/95% confidence intervals of the 1st-order coefficients from the 3rd-order polynomial curve fitting, per reference level, in the fixed ITD case.

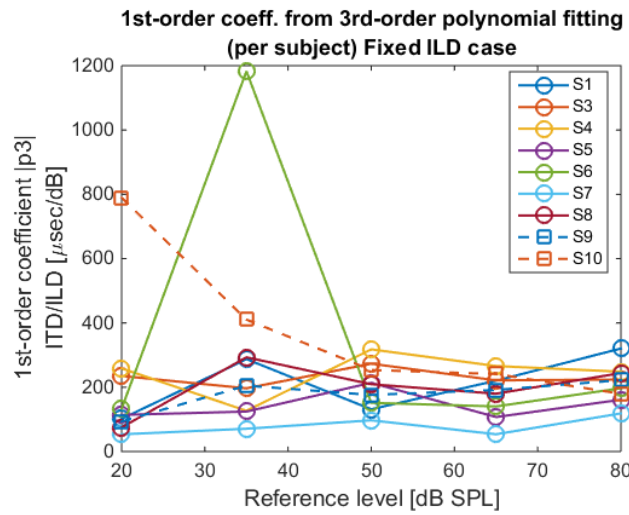


Figure 3.36: 1st-order coefficient from the 3rd-order polynomial curve fitting of individual dataset, per reference level, in the fixed ILD case.

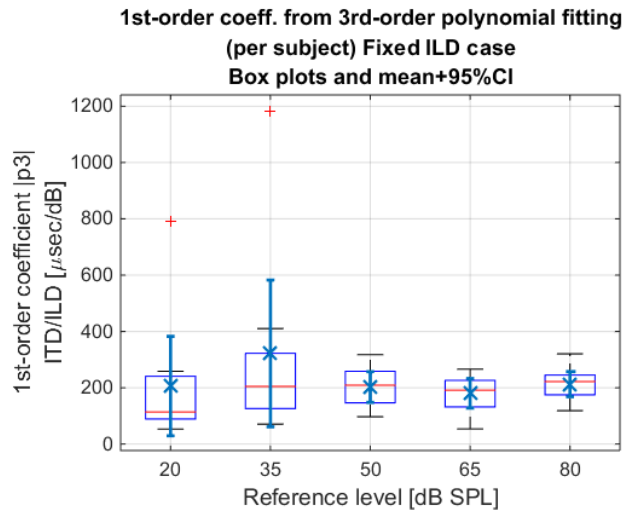


Figure 3.37: Box plots and mean/95% confidence intervals of the 1st-order coefficients from the 3rd-order polynomial curve fitting, per reference level, in the fixed ILD case.

the perceived strength of ILD increases. S6 has an outlying ratio at 35 dB reference level. Observation of the individual raw data showed that this subject could not make valid judgment in these cases, resulting in variable ITDs over 1800 μ s, which were removed as invalid data points for the analysis.

In both the fixed ITD and fixed ILD cases, independent-samples Kruskal-Wallis test showed that the null hypothesis (The 1st-order coefficient is the same across the reference levels) could not be rejected at .05 significance level. However, this step of analysis might not be meaningful after all because the coefficients were already the results of estimation through the polynomial fitting, sometimes with a very low R2 values. Instead, another hypothesis test was conducted, using the ratios directly derived for all (ILD-ITD) or (ITD-ILD) pairs. For example, in the fixed ITD case, the ratio ILD/ITD was calculated for the results from all the subjects, from the pairs where both the ILD and ITD were not zero. In the fixed ILD case, the ratio ITD/ILD was calculated in the same way. These ratios were used as the dependent variable for the same Kruskal-Wallis test, where the null hypothesis would be that the ratio is the same across the tested reference levels. In both the fixed ITD and fixed ILD cases, the null hypothesis was rejected ($p = .008$, and $.000$).

3.6.3 Discussion

The overall results imply that the reference level has some effect on the internally perceived strength of ILD. Despite the large variance, especially in the fixed ITD part, the tendency is such that at the lowest tested levels (20 dB and 35 dB) the ILD is perceived stronger as

the reference level increases, that for the reference levels from 35 dB to 65 dB the ILD is perceived weaker as the level increases, and that for the highest reference level of 80 dB the ILD is again perceived stronger as the level increases. This finding conforms to the finding of David *et al.* (1959), except at the 20 dB reference level. It also corresponds to the expectation from the computational simulation with the level-dependent basilar membrane nonlinearity introduced (by means of the use of the DRNL filterbank), whose example for a 1-kHz tone input is shown in Fig. 3.38. The internal ILD is decreased from the stimulus ILD at the nonlinear operating region, except at the high reference level where the nonlinearity is expected to be reduced.

Further investigations revealed that the large variance of the data in the current experiment, and the discrepancy between this experiment and that of David *et al.* (1959) may be because of the frequency of the tone. More specifically, it was found that with the tone frequency at 2 kHz, the ITD steps were not easily distinguishable regardless of the reference level, although the frequency had been chosen in the expectation that both the ITD and ILD would be effective in the lateralisation process. A lower frequency tone, or a different type of stimuli would have been more suitable to more clearly reveal the tendency. However, the overall level-dependent variations in the strength of ILD do seem to exist, and in this sense, the DRNL-based model in the TWO!EARS auditory front-end can work as a useful tool for further in-depth investigations into level-dependency of attributes of spatial auditory perception.

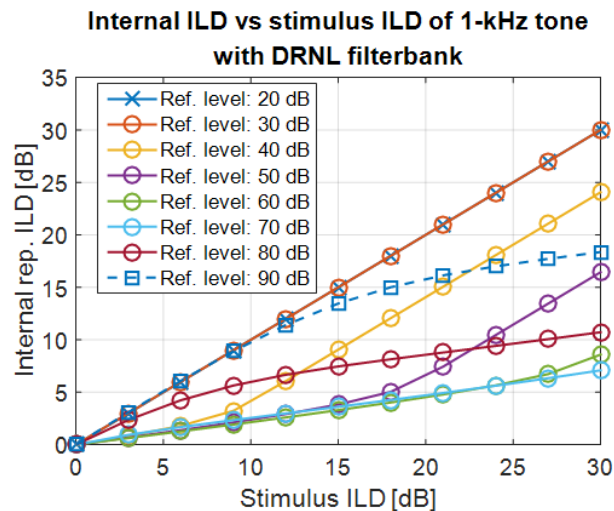


Figure 3.38: Simulation of ILD based on the internal representation of a 1-kHz tone over various reference levels, using the DRNL filterbank in the TWO!EARS auditory front-end.

4 Final quality model

In this chapter we describe the final model parts for the quality modeling.

4.1 Implementation of the van Dorp Schuitman (2011) model within the Two!Ears AFE framework

In Section 3.5, the so-called van Dorp Schuitman model was introduced and validated in a listening test, which predicts four acoustic attributes – reverberance, clarity, ASW and LEV – from binaural input, based on its internal representation. This modelling approach was unconventional compared to the physical property-based acoustical parameters, in that the internal representations at various peripheral stages are used for the prediction. This is also very relevant to the Two!EARS framework and its use-cases, although the van Dorp Schuitman model was basically devised and validated for the evaluation of room acoustics. This section introduces the van Dorp Schuitman model and its overall structure, and describes the attempt to integrate this model into the Two!EARS framework by implementing its algorithm using the existing AFE processors.

4.1.1 Model structure

Figure 4.1 shows the schematic structure of the processing in the van Dorp Schuitman model. Once the binaural input signal is given, it derives some monaural and binaural cues that are known to be related to the perception of the four acoustic attributes, and combines these internal representations in specific ways to predict them. More details of the individual processing stages can be found in van Dorp Schuitman (2011) and in van Dorp Schuitman *et al.* (2013).

The input signal firstly passes the outer/middle ear filtering process. Then, the gammatone-filterbank-based basilar membrane model transforms it into time-frequency domain representation. After this follows the inner hair cells model, which mimics the loss of high-frequency information except for the signal envelope. The next step is the application of frequency-dependent absolute threshold of hearing, followed by the adaptation loop, which mimics the adaptive nonlinear operation of the auditory nerve fibre and models the level-dependent

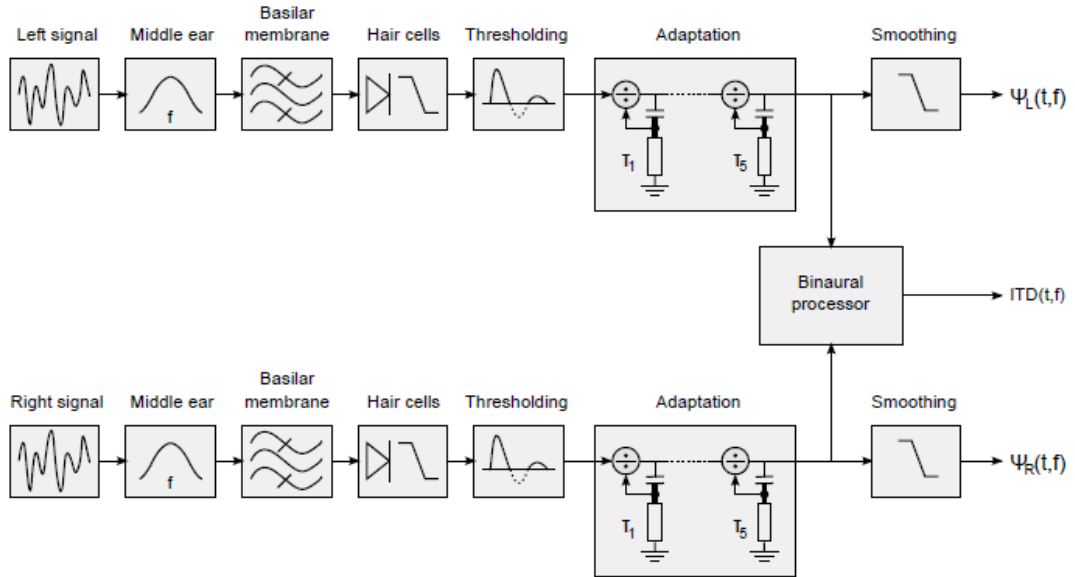


Figure 4.1: Processing structure of the van Dorp Schuitman model to derive the four acoustic properties (reproduced from van Dorp Schuitman (2011)).

masking effects. Most of these individual processing stages correspond well with the TWO!EARS auditory front-end processors; the outer- and middle-ear filtering, gammatone filterbank, inner hair cell, threshold-applied adaptation loop processors are already implemented within the TWO!EARS framework. Deliverable D 2.2 (Chapter 4) contains the detailed descriptions of these corresponding processors.

Based on the adaptation loop outputs, two types of inputs to the central processor are derived for the model output calculation: a further-smoothed monaural adaptation loop output (denoted by $\Psi_L(\tau, \mathbf{f})$ and $\Psi_R(\tau, \mathbf{f})$), and the ITD between the binaural adaptation loop outputs (ITD). The following processing happens in the central processor:

- The monaural inputs are divided into direct streams and reverberant streams, through an internal segregation algorithm based on thresholds and duration
- The ITD frames are also identified as the direct / reverberant parts, based on the corresponding segregated monaural inputs
- The direct/reverberant monaural streams and ITD frames are used in specific equations to derive the four raw outputs named $pCla$ (clarity), $pRev$ (reverberance), $pASW$ (apparent source width) and $pLEV$ (listener envelopment), which are then scaled to be between 0 and 1, named $sCla$, $sRev$, $sASW$ and $sLEV$, respectively

4.1.2 Implementation of model

As mentioned earlier, most of the peripheral processing steps involved in the van Dorp Schuitman model correspond to the already available TWO!EARS auditory front-end processors. Therefore, the implementation of the model and its integration was carried out by using the corresponding auditory front-end chain, and by following the descriptions in the work of van Dorp Schuitman (2011).

Here, a number of internal constants or parameters were introduced for calculation of the outputs from the monaural streams and the ITD frames. An example is the set of parameters used for the direct/reverberant stream segregation – peak and dip thresholds, and the duration threshold determine which portions of the input streams can be regarded as the direct or reverberant parts. These internal parameters were optimised via a genetic algorithm such that the model outputs would match the responses from listeners in a set of subjective tests using the same binaural stimuli. During the initial implementation and validation stages, it was found that the same values of these parameters in the TWO!EARS version of the model did not result in the same output values of the 4 attributes as with the original model. This can be due to a) slight mismatches in the internal parameters within each of the sub-processors (despite the efforts to individually match the outputs at corresponding stages as closely as possible), and b) some unknown updates or changes in the original model implemented after the publications. Moreover, this mismatch also implies that although the original model had been optimised and validated against a set of experimental results, it may not guarantee that the outputs will always be in a reasonable range or order with a new set of stimuli. Therefore, a decision was made to provide the key internal parameters as a structured input to the model, as well as the binaural stimulus. This way, it would be possible to fine-tune the model as necessary, depending on the circumstances, through any optimisation technique.

As a result, a function `afeRAA.m` was implemented which takes a binaural `.wav` file and a parameter structure as the input, and calculates the output structures as the original model. The following list describes the main input configuration parameters defined as a set named `parConf` in the TWO!EARS van Dorp Schuitman model, that can be optimised as in the original model depending as needed:

- `parConf.mu_psi`: Relative level of the peak threshold (for direct/reverberant stream segregation) (μ_{Ψ} in van Dorp Schuitman (2011))
- `parConf.mu_psi_dip`: Relative level of the dip threshold ($\mu_{\Psi,dip}$ in van Dorp Schuitman (2011))
- `parConf.T_min`: Minimum width of a peak or dip (T_{min} in van Dorp Schuitman (2011))

- `parConf.mu_ASW`: Weighting factor used in ASW calculation (μ_{ASW} in van Dorp Schuitman (2011))
- `parConf.nu_ASW`: Weighting factor used in ASW calculation (ν_{ASW} in van Dorp Schuitman (2011))
- `parConf.mu_LEV`: Weighting factor used in LEV calculation (μ_{LEV} in van Dorp Schuitman (2011))
- `parConf.nu_LEV`: Weighting factor used in LEV calculation (ν_{LEV} in van Dorp Schuitman (2011))

And the following list describes the key output structure named `par`:

- `par.FL`: Level of foreground stream (in Model Units (MU), as from the adaptation loop processor)
- `par.BL`: Level of background stream (MU)
- `par.ITDf`: ITD fluctuation in the foreground (sec)
- `par.ITDb`: ITD fluctuation in the background (sec)
- `par.Llow`: Level of the low-frequency part of the spectrum
- `par.pRev`: Reverberance (unscaled)
- `par.pClar`: Clarity (unscaled)
- `par.pASW`: Apparent Source Width (unscaled)
- `par.pLEV`: Listener Envelopment (unscaled)
- `par.sRev`: Reverberance (scaled)
- `par.sClar`: Clarity (scaled)
- `par.sASW`: Apparent Source Width (scaled)
- `par.sLEV`: Listener Envelopment (scaled)

More detailed descriptions of how these input/output parameters are defined and used can be found in the two publications van Dorp Schuitman (2011) and van Dorp Schuitman *et al.* (2013).

4.1.3 Discussion and outlook

The van Dorp Schuitman model can be regarded as an initial step out of the conventional prediction techniques of room acoustics that mainly depend on physical measurements and are being questioned recently (Blauert and Raake, 2015). Be it the room acoustics evaluation, or the spatial audio quality of experience evaluation, it is now increasingly argued that these evaluation processes involve very complicated interaction between the peripheral and cognitive listening stages. The van Dorp Schuitman model, utilising some of the more complex and non-linear peripheral processors, was therefore worth integrating into the TWO!EARS system. This will open up many possibilities for future research, in that now this perceptual quality prediction model can be simulated/refined further on its own through optimisation, or be investigated in conjunction with some of the higher cognitive processing stages.

4.2 Preference prediction

4.2.1 Motivation and modeling goal

Towards modeling the Quality of Experience of spatial audio reproduction systems, this section presents a first study on the algorithmic prediction of preference ratings. Being in line with the considerations in Chapter 2 and the corresponding experimental paradigm of paired comparisons in Chapter 3, we decided to focus here on the specific task of directly predicting the pairwise comparisons between stimuli. The motivation was that the human listeners had the same task of rating pairwise preferences, and giving the same task to the model is not only *fair* to the model, but also more reliable since the model uses directly those perceptual ratings as ground-truth, and not processed preference scores such as the Bradley-Terry-Luce scores.

Considering which of the different perceptual experiments in Chapter 3 the model should aim to predict, the choice fell on Experiment 3 of the series on comparing different spatial audio reproduction systems (Section 3.1). First, the experiment on easier scenes (Section 3.4) did not reveal clear insights on the aspects contributing to preference (except the impact of knowing where it is actually positioned) and the colouration experiment (Section 3.3) showed only that colouration is affecting quality, but did not yield interpretable dimensions that allowed to identify individual aspects of colouration that contribute to preference. Second, the experiments presented in Section 3.1 gave a more complete picture on the aspects contributing to sound quality (preferences) of spatial reproduction systems, as they included the interaction of mixing techniques and the audio reproduction system. Finally, it also verified that dynamic binaural synthesis can be used to simulate the involved spatial audio systems, meaning it used binaural stimuli that can directly be fed into a binaural

auditory model, such as for instance the different processors of the TWO!EARS Auditory Frontend.

With the choice of experiment 3 from Section 3.1 as ground-truth data, the model purpose needs to be clarified given the type of conditions used in that experiment. Except for the stereo condition, all stimuli concerned different mixing parameters in conjunction with a WFS system. Thus, training and evaluating a model on this data means the model purpose is actually not predominantly the preference for different spatial audio reproduction systems, but rather the preference for different mixing parameters, which also explains the title of this study.

It is noted that this study is intended to show a first proof-of-concept of such a model. It is not claimed to represent a robust model that can directly be applied to different experimental contexts. The reason for such a disclaimer lies in the fundamental limitation of the ground-truth data: only one music piece was available, hence the validity of the model for other music will need to be investigated in future research.

Given this proof-of-concept approach, this section does not only discuss the achieved modelling results; it also puts a strong weight on reporting the details on how to build such a model. Furthermore, it is noted that the discussion of results is focussed on the predictive power of the different features used, and less on the performance in absolute numbers as provided as well to show the successful proof-of-concept.

4.2.2 Perceptual data

Experimental data and definition of comparison groups For the present work, the perceptual data from experiment 3 on comparing different spatial audio reproduction systems (Section 3.1) was used. It is noted that the WFS stimuli used in that experiment differed in four mixing parameters: compression, equalization, reverb, and positioning. In addition, a fifth mixing parameter, “vocal processing” was also applied, in which the previously mentioned mixing parameters were modified only on the vocals. There were three different instantiations for the first three mixing parameters and four for the latter two. Together with the WFS reference mix and a stereo version, this resulted in 19 stimuli, as summarized in Table 4.1.

The experimental design was such that test participants had to compare either two stimuli that represented two alternatives for the same mixing parameter, e.g. COMP_M and COMP_MM, or two stimuli that differed in the mixing parameters, e.g. COMP_M and EQ_MM. In that respect, it is possible to organize the pairwise combinations of the 19 conditions in six *comparison groups*: one group for each mixing parameter in which only stimuli of that mixing parameter were used in a comparison (*Compression, EQ, Reverb, Positioning, Vocals*), and one additional group *Mixed* for comparisons between stimuli

| Name | Mixing parameter | Short description |
|---------|------------------|---|
| WFS | — | WFS reference mix |
| STE | — | Stereo mix |
| COMP_M | Compression | Half of the gain reduction of the compressor, compared to the WFS reference mix |
| COMP_MM | | Compressor switched off |
| COMP_P | | Double of the gain reduction of the compressor , compared to the WFS reference mix |
| EQ_M | Equalization | Half of the amount of cutting or boosting in the EQ filters, compared to WFS mix |
| EQ_MM | | EQ filters switched off |
| EQ_P | | Double of the amount of cutting or boosting in the EQ filters, compared to WFS mix |
| REV_M | Reverb | Half of all reverb return signals, compared to WFS mix |
| REV_MM | | All return signals switched off |
| REV_P | | Double of all reverb return signals, compared to WFS mix |
| POS_M | Positioning | Shifting of foreground elements towards the center: wider than stereo, narrower than WFS |
| POS_MM | | Shifting of foreground elements towards the center: within the two stereo speaker positions |
| POS_P | | Spreading of foreground elements farther apart and away from the center: slight spreading |
| POS_PP | | Spreading of foreground elements farther apart and away from the center: extreme spreading |
| VOC_M | Vocals | Reducing processing of vocals: half of Compression, EQ, and Reverb settings compared to WFS mix |
| VOC_MM | | Processing of vocals switched off |
| VOC_P | | Increasing processing of vocals: double of Compression, EQ, and Reverb settings compared to WFS mix |
| VOC_Pb | | Increasing processing of vocals, variant: half of Compression and EQ settings compared to WFS mix |

Table 4.1: List of the 19 stimuli used in the perceptual experiment.

that were modified in different mixing parameters. Comparisons including either a WFS reference or a stereo mix were assigned to the group of the other non-WFS or non-stereo stimulus; comparisons between the WFS and stereo mix were assigned to the mixed group. Note that this grouping has actually been used in the experimental design, in which test participants judged all comparisons per group in a row before continuing with the next group, whereas the order within each comparison group, as well as the order of the whole comparison groups was randomized across test participants.

Since the later modeling approach as well as the discussions of results will heavily rely on those comparison groups, Table 4.2 lists all combinations assigned to the different comparison groups.

4 Final quality model

| Compression | Equalization | Reverb | |
|--------------------|----------------|------------------|--|
| WFS vs. COMP_M | WFS vs. EQ_M | WFS vs. REV_M | |
| WFS vs. COMP_MM | WFS vs. EQ_MM | WFS vs. REV_MM | |
| WFS vs. COMP_P | WFS vs. EQ_P | WFS vs. REV_P | |
| STE vs. COMP_M | STE vs. EQ_M | STE vs. REV_M | |
| STE vs. COMP_MM | STE vs. EQ_MM | STE vs. REV_MM | |
| STE vs. COMP_P | STE vs. EQ_P | STE vs. REV_P | |
| COMP_M vs. COMP_MM | EQ_M vs. EQ_MM | REV_M vs. REV_MM | |
| COMP_M vs. COMP_P | EQ_M vs. EQ_P | REV_M vs. REV_P | |
| COMP_MM vs. COMP_P | EQ_MM vs. EQ_P | REV_MM vs. REV_P | |

| Positioning | Vocals | Mixed | |
|-------------------|-------------------|-------------------|---|
| WFS vs. POS_M | WFS vs. VOC_M | WFS vs. STE | |
| WFS vs. POS_MM | WFS vs. VOC_MM | COMP_x vs. EQ_y, | with x,y ∈ (M, MM, P) |
| WFS vs. POS_P | WFS vs. VOC_P | COMP_x vs. REV_y, | with x,y ∈ (M, MM, P) |
| WFS vs. POS_PP | WFS vs. VOC_PP | COMP_x vs. POS_y, | with x ∈ (M, MM, P), y in (M, MM, P, PP) |
| STE vs. POS_M | STE vs. VOC_M | | |
| STE vs. POS_MM | STE vs. VOC_MM | COMP_x vs. VOC_y, | with x ∈ (M, MM, P), y in (M, MM, P, Pb) |
| STE vs. POS_P | STE vs. VOC_P | | |
| STE vs. POS_PP | STE vs. VOC_PP | EQ_x vs. REV_y, | with x,y ∈ (M, MM, P) |
| POS_M vs. POS_MM | VOC_M vs. VOC_MM | EQ_x vs. POS_y, | with x ∈ (M, MM, P), y in (M, MM, P, PP) |
| POS_M vs. POS_P | VOC_M vs. VOC_P | | |
| POS_M vs. POS_PP | VOC_M vs. VOC_Pb | EQ_x vs. VOC_y, | with x ∈ (M, MM, P), y in (M, MM, P, Pb) |
| POS_MM vs. POS_P | VOC_MM vs. VOC_P | | |
| POS_MM vs. POS_PP | VOC_MM vs. VOC_Pb | REV_x vs. POS_y, | with x ∈ (M, MM, P), y in (M, MM, P, PP) |
| POS_P vs. POS_PP | VOC_P vs. VOC_Pb | REV_x vs. VOC_y, | with x ∈ (M, MM, P), y in (M, MM, P, Pb) |
| | | POS_x vs. VOC_y, | with x ∈ (M, MM, P, PP), y in (M, MM, P, Pb) |

Table 4.2: Overview of comparison groups, defining which stimuli pairs belong to which group. Note: 48 of the 116 possible combinations ($\approx 41\%$) for the Mixed group were actually tested.

Computing preference matrices and preference probabilities In order to prepare the raw perceptual data for modeling, two processing steps were conducted. First, the data for the individual pairwise comparisons between two stimuli was combined into preference matrices. A preference matrix *prefMat* contains the counts how often a stimulus in row *a* was preferred over another stimulus in column *b*. This computation was done separately for each of the six comparison groups, as well as for the whole data set combined. As an example, Table 4.3 shows the preference matrix for the comparison group *Compression*.

Second, the counts in the preference matrices can be transformed into a probability value $p(a \succ b)$ indicating the likelihood that a stimulus *a* is preferred over a stimulus *b*. This can

| | WFS | STE | COMP_M | COMP_MM | COMP_P |
|---------|-----|-----|--------|---------|--------|
| WFS | - | 30 | 20 | 30 | 27 |
| STE | 11 | - | 8 | 18 | 11 |
| COMP_M | 21 | 33 | - | 23 | 21 |
| COMP_MM | 11 | 23 | 18 | - | 18 |
| COMP_P | 14 | 30 | 20 | 23 | - |

Table 4.3: Example: Preference matrix per the group *Compression*. Read each matrix element as follows: stimulus in row a is x -times preferred over stimulus in column b .

be computed from the preference matrix $prefMat$ by the equation

$$p(a \succ b) = \frac{prefMat(a, b)}{prefMat(a, b) + prefMat(b, a)}$$

As an example, Table 4.4 shows the preference probabilities for the *Compression* group.

Discussion One important observation needs to be made for the perceptual data: some stimulus pairs do not show a clear preference rating across test participants. On the one hand, this is already visible in the preference matrix, since the numbers in the matrix elements $prefMat(a, b)$ are often similar to the numbers in the corresponding *inverse* matrix elements $prefMat(b, a)$. On the other hand, this is clearly visible in the probability tables, where many combinations show a probability close to 0.5.

This has some implications for the modeling: From a practical perspective, in terms of treating those stimulus pairs with such unclear preference, and from a conceptual perspective in terms of interpreting modeling performance in view of such cases in the ground-truth data. These will be addressed by investigating the model performance for different subsets of the data, including or excluding such unclear cases.

| | STE | COMP_M | COMP_MM | COMP_P |
|---------|------|--------|---------|--------|
| WFS | 0.73 | 0.49 | 0.73 | 0.66 |
| STE | | 0.20 | 0.44 | 0.27 |
| COMP_M | | | 0.56 | 0.51 |
| COMP_MM | | | | 0.44 |

Table 4.4: Example: Preference probabilities for the group *Compression*. Read each table element as follows: the probability $p(a \succ b)$ that stimulus in row a is preferred over stimulus in column b is the table entry x . The “inverse” interpretation $p(b \succ a)$, i.e. stimulus in column b is preferred over stimulus in row a is $1 - x$.

4.2.3 Modeling approach

The general model structure consists of three main steps: first, compute feature vector representations of each stimulus F_i ; then, compute a difference of the feature vectors of each pair of stimuli $D_{i,j}$; and finally, map such a difference vector on the perceptual ground-truth preference data $P(i \succ j)$ in order to predict which of the corresponding two stimuli is preferred $\hat{P}(i \succ j)$.

Implications on data processing when computing the difference feature vectors This modelling approach requires a computation of a difference between feature vectors that keeps the sign, i.e. $D_{a,b} = F_a - F_b$. Otherwise a conventional distance computation ignoring the sign, such as $D_{a,b} = |F_a - F_b|$, would not allow an identification which of the two feature vectors would be more preferred. This requirement of keeping the sign, however, implies that the order of stimuli put into the model would influence the result, as $D_{a,b} \neq D_{b,a}$. To deal with this, the model needs to be able to predict which stimulus is preferred for both cases $D_{a,b}$ and $D_{b,a}$, as the model cannot know upfront which stimulus is sent to which input a or b .

As an example, suppose a stimulus X is preferred over a stimulus Y . Then, if the model gets X into its input a , and Y into its input b , then the model evaluates $D_{a,b} = D_{X,Y}$ and must predict “ X better Y , i.e. input a better than input b ”. Assuming the model assigns a positive preference value $\hat{P}(a,b)$ if the stimulus in input a is indeed more preferred than the stimulus in input b , and a negative $\hat{P}(a,b)$ for the opposite case, then this means, for the current example $\hat{P}(X,Y) = \text{model}(D_{a,b})|_{a=X,b=Y} > 0$. However, if the model now gets Y as its input a , and X as its input b , then the model evaluates $D_{a,b} = D_{Y,X}$. Since it still must predict “ X better Y , i.e. now input a worse than input b ”, this means $\hat{P}(Y,X) = \text{model}(D_{a,b})|_{a=Y,b=X} < 0$.

The way to achieve such behavior is to confront (i.e. train and test) the model with both cases, i.e. the difference feature vectors $D_{X,Y}$ with the corresponding ground-truth preference values $P(X,Y) > 0$ and the *inverted* difference feature vectors $D_{Y,X}$ with the corresponding *inverted* ground-truth values $P(Y,X) < 0$.

Computation of target values for the model output The modeling approach suggests that the model is essentially a two-class classifier, which is trained and tested against ground-truth data, i.e. the perceptual preference data. Thus, the straight-forward approach is to assign as target values for each stimulus pair a class label of -1 or $+1$, depending on which stimulus is preferred according to the preference matrices of Section 4.2.2. Alternatively, one can also use the preference probability values of Section 4.2.2 and interpret them as class probabilities.

Classification method Considering the model task as a two-class problem, Support Vector Machines (SVM) are the primary choice, since SVM belong to the category of two-class classifiers, and they have been successfully used for audio and music classification problems for at least a decade International Society for Music Information Retrieval (2016).

Considerations on the issue of over-training It is known in the field that limited data sets can lead to overtraining effects. Generally speaking, over-training means the classifier is able to reproduce the available data, but not the actual structure behind that data. As a result, the classifier performance on the available but limited data is very high, while it is likely that the classifier performance on new data will be much worse.

There are two aspects concerning such a limited generalizability: how well the limited data represents the problem domain, and how many features may be used before over-training effects occur. Concerning the first aspect, only one music piece was used for this study. To develop a general preference prediction model for a variety of musical pieces is a complex undertaking that requires future research solely dedicated to this task.

Concerning the second aspect, one typical approach is to limit the number of features to a minimum, which still allows for good classification performance. While Support Vector Machines (SVMs) are designed to use high-dimensional data, i.e. large feature numbers, empirical evidence was found (e.g. Chen and Lin (2006), Pal and Foody (2010)) that also SVM can benefit from feature selection. With these considerations in mind, we decided to limit the number of features in order to minimize over-training effects given the rather small data set of 206 data points (i.e. nine comparisons for the groups *Compression*, *EQ*, and *Reverb*, 14 for the groups *Positioning* and *Vocals*, 151 for the group *Mixed*).

4.2.4 Feature extraction

Since four different mixing parameters (compression, equalization, reverb, positioning) were changed to generate the different stimuli, we hypothesized that a potentially powerful feature set should comprise four different feature types, each dedicated to characterize one mixing parameter. Due to the considerations on over-training (see above), we aimed at a maximum of four features per feature type, which is about half the number of data points for the smallest comparison groups (i.e. nine for *Compression*, *EQ*, and *Reverb*). Table 4.5 provides an overview of the 15 chosen features, while the following text provides a more detailed description and motivation for their inclusion into this study.

Features characterizing Compression Skovenborg (2014) evaluated a number of typical *signal peak to signal average* measures and found that they hardly correlate with perceived dynamics. For that reason, he proposed another feature called *LDR* that describes microdynamic behavior in music by comparing a *fast* loudness function using a short integration time with a *slow* loudness function using a long integration time. Since this feature showed a better correlation than the other measures Skovenborg tested, we chose this feature as a candidate for characterizing compression. More precisely, we adopted the computation of the LDR feature for the TWO!EARS framework by computing the slow and fast loudness signal from the gammatone filterbank outputs of the Auditory Frontend. In addition, we decided to further exploit the concept behind LDR by computing two measures that also characterize the relation between fast and slow loudness signals: the Pearson correlation coefficient and the Root Mean Square Error.

Features characterizing Equalization Since Equalization means to modify the relative amount of signal energy in different frequency areas, typical spectral features such as spectral centroid or spectral flatness are good candidates. In his master’s thesis, Nagel (2016) investigated the potential usefulness of spectral features to characterize the Bradley-Terry-Luce scores of a first subset of the perceptual data (21 test participants instead of 41). Nagel investigated six spectral features, all computed with the TWO!EARS Auditory Frontend from the Rate Map, which means those features characterize the spectral energy distribution in a perceptual domain and not in the signal frequency domain. For the current work we took the four most promising features from the six available ones: *Decrease*, *variation*, *entropy*, and *irregularity*. The spectral *decrease* serves to investigate differences in the low frequency information in the signals, and is computed from the binaural signals Peeters *et al.* (2011). This is done by examining the spectral slope of the rate-map, with an emphasis on the low frequencies. The *entropy* and *irregularity* are calculated from the rate-map of the signals. The entropy measures the peakiness of the signal, resulting in a low value for rate-maps with many distinct spectral peaks and a high value for flatter spectra Misra *et al.* (2005). To further investigate how the rate-map changes with time, the ‘variation’ of the rate-map was extracted. This is defined as one minus the correlation between two adjacent time-frames of the rate-map.

Features characterizing Reverb van Dorp Schuitman *et al.* (2013) developed a model to estimate room acoustic parameters from binaural input signals, as described in the previous Section 4.1. This model computes a perceptual representation of the input signal, splits this signal into a foreground and background stream, and computes four room acoustic parameters. Those parameters are (with quotations from the original authors): *reverberance* (“amount of reverberation perceived by listeners”), *clarity* (“degree to which discrete sounds in a signal stand apart in time from one another subjectively”), *apparent source width ASW* (“apparent broadening of a sound source can occur as a result of early

lateral reflections, resulting in a certain ASW”), and *listener envelopment* *LEV* (“perceptual parameter related to spaciousness and refers to the environment instead to the source”). Hypothesizing that applying reverb effects in a music mix leads to a similar perception as the perception of the acoustical properties of a real room, and considering that the van Dorp Schuitman model just needs binaural recordings, we considered this model as a good candidate for the characterization of the *reverb* mixing parameter. While the reverberance parameter of the van Dorp Schuitman model is an obvious candidate, we decided to also include the other three parameters, as we hypothesized that applying reverb as a mixing parameter can also affect clarity (e.g. by temporal smearing of musical events), ASW (e.g. for background sounds in the music mix) and LEV (e.g. by increasing impression that the reverb comes from many directions). Since it was started before the Van Dorp Schuitman model has been integrated into the TWO!EARS framework, we used the original implementation for this modelling at first.

Features characterizing Positioning Hearing different music elements in a mix at different positions means that they have different localization cues. Obviously, ITD and ILD are then primary candidate features for characterizing Positioning. For the present model, we decided to directly exploit the ITD and ILD cues extracted from the TWO!EARS Auditory Frontend. Motivated by the hypothesis that the different Positioning mixes lead to different variations of ITD and ILD cues over time and frequency, we computed for ITD and ILD as features two of the four combinations of the mean and standard deviation (STD) operation across time and frequency (Auditory Frontend bands): Mean over bands & STD over time, STD over bands & STD over time.

4.2.5 Model variations using different feature subsets

Since the purpose of this study is to proof the concept of modeling preferences, it is reasonable to investigate in more detail the *power* of the respective four different feature types. For that reason we decided to test a number of models using different combinations of the features. The idea is to compare models in which either all four feature types are used, in which only one feature types is used, and in which all except one feature types is used. This kind of cross-comparison allows to interpret on the importance of each feature type. Thus, nine models according to Table 4.6 are evaluated in this work.

4.2.6 Model evaluation

Training and validation method Since the data set is rather small, we opted for the a mixed cross-validation / bootstrap method for model evaluation, the latter being developed by Efron (1992) to estimate the true value of the model performance for small data sets.

4 Final quality model

| Feature Type | No. | Name / Description | Target Mixing Parameter |
|--------------|-----|--|-------------------------|
| LDR | 1 | LDR_Diff Variation of Skovenborg's LDR Skovenborg (2014): 95 percentile of difference (in dB) between slow loudness (3s integration time) and fast loudness (25ms), using gammatone-filter output of Two!EARS auditory front-end, averaging across filters by arithmetic mean | Compression |
| | 2 | LDR_Rho Addition based on Skovenborg's LDR Skovenborg (2014): Pearson correlation coefficient between slow and fast loudness, averaging across filters by arithmetic mean | |
| | 3 | LDR_RMSE Alternative to LDR_Diff: Root Mean Square Error between slow and fast loudness | |
| SPEC | 4 | Decrease Two!EARS Auditory Frond-End (AFE) - Spectral Features Processor: average spectral slope of the rate-map representation, putting a stronger emphasis on the low frequencies | Equalization |
| | 5 | Variation AFE - Spectral Features Processor: defined as one minus the normalised correlation between two adjacent time frames of the rate-map | |
| | 6 | Entropy AFE - Spectral Features Processor: peakiness of the spectral representation, low for a rate-map with many distinct spectral peaks and high for a flat rate-map spectrum | |
| | 7 | Irregularity AFE - Spectral Features Processor: quantifies the variations of the logarithmically-scaled rate-map across frequencies | |
| LOC | 8 | ITD_stdTime_meanFilters post-processing of AFE - ITD Processor: variation of ITDs over time, on average across filterbank channels | Positioning |
| | 9 | ITD_stdTime_stdFilters post-processing of AFE - ITD Processor: variation of ITDs over time and across filterbank channels | |
| | 10 | ILD_stdTime_meanFilters post-processing of AFE - ILD Processor: variation of ILDs over time, on average across filterbank channels | |
| | 1 | ILD_stdTime_stdFilters post-processing of AFE - ILD Processor: variation of ILDs over time and across filterbank channels | |
| VDS | 12 | sRev (scaled Reverb) Reverb feature of vanDorpSchuitman (vDS) model, using original implementation , using level-normalized input signals | Reverb |
| | 1 | sClar (scaled Clarity) Clarity feature of vDS model, using original implementation , using level-normalized input signals | |
| | 14 | sASW (scaled Apparent Source Width) Apparent Source Width feature of vDS model, using original implementation , using level-normalized input signals | |
| | 15 | sLEV (scaled Listener Envelopment) Listener Envelopment feature of vDS model, using original implementation , using level-normalized input signals | |
| | | | |

Table 4.5: Overview of the 15 features assigned to the four chosen feature categories.

| Model | Feature types included | | | |
|-------|------------------------|------|-----|-----|
| | LDR | SPEC | VDS | LOC |
| M1 | ✓ | ✓ | ✓ | ✓ |
| M2 | ✓ | | | |
| M3 | | ✓ | | |
| M4 | | | ✓ | |
| M5 | | | | ✓ |
| M6 | | ✓ | ✓ | ✓ |
| M7 | ✓ | | ✓ | ✓ |
| M8 | ✓ | ✓ | | ✓ |
| M9 | ✓ | ✓ | ✓ | |

Table 4.6: Overview of the nine tested models resulting from different combinations of the feature sets used.

The principle of the implemented approach is to conduct many repetitions (typically in the order of 100 to 1000), in which the data is randomly split into training or test sets, whereas the split is – in contrast to k-fold cross validation or leave-one-out method – done independently from the previous iteration. We opted to run 100 bootstrap repetitions with a training to test split of 80-to-20.

Using an algorithm developed for another modeling task (Skowronek, 2016), the training and test split was done such that the algorithm attempted to achieve this target ration per comparison group. The motivation is to minimize the occurrence of bootstrap repetitions in which no data for an individual comparison group is selected as test data.

Performance measures The conventional way to asses the performance of a classifier is to compute a confusion matrix which shows how many percent of data points are assigned to the correct class or to the other – incorrect – classes. The overall performance is then determined by the percentages on the main diagonal of such a confusion matrix, often by computing the mean across those main diagonal elements.

The fact that we conducted bootstrap-type repetitions allows to compute average performance by computing the mean of the confusion matrix elements across repetitions, leading to a *Mean Confusion Matrix*. And it also allows to compute a measure of the model stability in terms of the performance variability across repetitions by computing the 95% confidence intervals (CI95) of the confusion matrix elements across repetitions, leading to a *CI95 Confusion Matrix*.

Bearing this in mind, the model performance will be reported as the mean across the (here) two main diagonal elements of the *Mean Confusion Matrix* and the *CI95 Confusion Matrix*, expressing the average performance and the *average* stability.

Subsets of ground-truth data There are two aspects that allow to define different subsets of the ground-truth data, i.e. the perceptual preference values. First, one can select either only data containing comparisons within a group (in terms of the mixing parameter), or all data including also comparisons of stimuli across groups. Second, one can select either only data for which the perceptual ground-truth shows a clear preference, or all data including also cases with an unclear preference. Here, a *clear preference* is defined as preferences beyond a certain threshold in terms of the preference probability: if the preference probability (for the first stimulus in a comparison) lies between 0.4 and 0.6 (i.e., 0.1 around the chance level of 0.5), then it is not clear whether this first stimulus is preferred over the second (preference probability > 0.6) or vice versa (preference probability < 0.4).

The resulting four data sets and their benefits are:

- **Data Set 1:** Within-group comparisons only, comparisons only with clear preferences. This is the easiest data set. It allows to check the basic potential of the four different feature sets and it is a *fair* task for the model, as it is compared only against ground-truth data in which also humans show a clear preference. There are 50 data points in this data set.
- **Data Set 2:** Within-group comparisons only, comparisons with clear and unclear preferences. This data set still allows to discuss the basic potential of the four different feature sets, but it enables to discuss the sensitivity of the approach in case of *noisy* data in terms of the clearness of human preferences. There are 110 data points in this data set.
- **Data Set 3:** Within-group and across-group comparisons, only comparison with clear preferences. This data set allows to investigate the appropriateness of the four feature sets when confronted with comparisons of stimuli that differ in multiple mixing parameters. Still it is a *fair* task, as it focusses on the clear ground-truth comparisons. There are 66 data points in this data set.
- **Data Set 4:** Within-group and across-group comparisons, comparisons with clear and unclear preferences. This is the most difficult but also most realistic data set for modeling. It provides good insights on the feasibility for real-life scenarios. There are 206 data points in

this data set.

Evaluation on whole data or per comparison group The evaluation algorithm allows to compute the performance across the whole data sets as well as per comparison group. The first option allows to assess the general model performance, the latter provides more diagnostic insights on how well the different mixing parameters (which define the groups) can be modelled. Both results will be reported.

4.2.7 Modeling performance

Overall performance and impact of different ground-truth data subsets Figure 4.2 shows the model performance for the four different data subsets for the nine tested models. The top left plot shows a performance of close to 100% for Data Set 1, which is much higher than the performance values for the other three data sets. Apparently this data set is very ideal and too small for reasonable classification, as this result is most likely showing overtraining effects, despite the small amount of features used. Thus we will not consider this Data Set anymore in the following analyses.

The other plots show more realistic performances for Data Sets 2, 3 and 4. As expected, the performance is – on average – decreasing from Data Set 2 to 4. A positive result is the fact that the model performance is in all cases above chance, in several cases (for Data Sets 2 and 3) even above 75%. This is a first observation showing that it is feasible to predict the perceptual preference data. More insightfull, however, is to discuss the benefit of the different feature types by investigating how the individual models relate to each other and how these relations are affected by the different data sets.

First we discuss the LDR features. The LDR features alone (Model M2) show the worst performance in all three data sets. Removing those features from the full model (M6) is similar to or even slightly better than keeping the full model (M1), depending on the data set. Thus, the LDR features do not clearly contribute to the model performance, i.e. they appear not to be useful for predicting the paired comparisons.

Next we discuss the SPEC features. The SPEC features alone (M3) show slightly lower performance for Data Sets 2 and 3 than in Data Set 4. Removing those features from the full model (M7) is slightly better than keeping the full model (M1) for Data Sets 2 and 3. For Data Set 4, however, removing those features causes a significant performance drop compared to the full model. Thus, the contribution of the SPEC features to the model performance appears to be sensitive to the data set used: in rather clean cases (Data Sets 2 and 3) the contribution is minor, probably even negative; but in the most challenging data set the contribution is significantly positive.

Next we discuss the VDS features. The VDS features alone (M4) show essentially the same performance as the full model (M1) for Data Set 2, a slightly better performance for Data Set 3, but a significantly worse performance for Data Set 4. Removing those features (M8), interestingly, shows essentially no performance change compared to the full model for all three Data Sets. Thus, the contribution of the VDS features to the model performance appears to be even more sensitive to the data set than the SPEC features.

Finally, the LOC features show a very similar behavior as the VDS features, and the same conclusions can be drawn.

To summarize, the LDR features appear to be dispensible, while the other feature types (SPEC, VDS and LOC) appear to be required, whereas dropping one of those three feature types can be compensated by the other two. However, these results are based on the performance across all six Comparison Groups. Since the different feature sets were intended to specifically address four different Comparison Groups, the next paragraph will investigate, in how far the different feature types can predict the individual comparison groups.

Performance per comparison group Since the different feature types are targeting the different mixing parameters, it is possible to formulate upfront a number of hypotheses, which of the different models can be expected to work well for the different comparison groups. Table 4.7 provides an overview of those hypotheses.

An important aspect in the following is the question of how to define a threshold for good and bad performance. Since the goal here is to obtain insights into the benefit

| Model | Feature types included | | | | Hypotheses: Good (✓) or bad (×) performance for groups... | | | | | |
|-------|------------------------|------|-----|-----|---|----|--------|-------------|--------|-------|
| | LDR | SPEC | VDS | LOC | Compression | EQ | Reverb | Positioning | Vocals | Mixed |
| M1 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| M2 | ✓ | | | | ✓ | | | | | |
| M3 | | ✓ | | | | ✓ | | | | |
| M4 | | | ✓ | | | | ✓ | | | |
| M5 | | | | ✓ | | | | ✓ | | |
| M6 | | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | | |
| M7 | ✓ | | ✓ | ✓ | ✓ | × | ✓ | ✓ | | |
| M8 | ✓ | ✓ | | ✓ | ✓ | ✓ | × | ✓ | | |
| M9 | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | × | | |

Table 4.7: Overview of the different hypotheses in terms of expected model performance for the different comparison groups, based on the different feature types included in the different models.

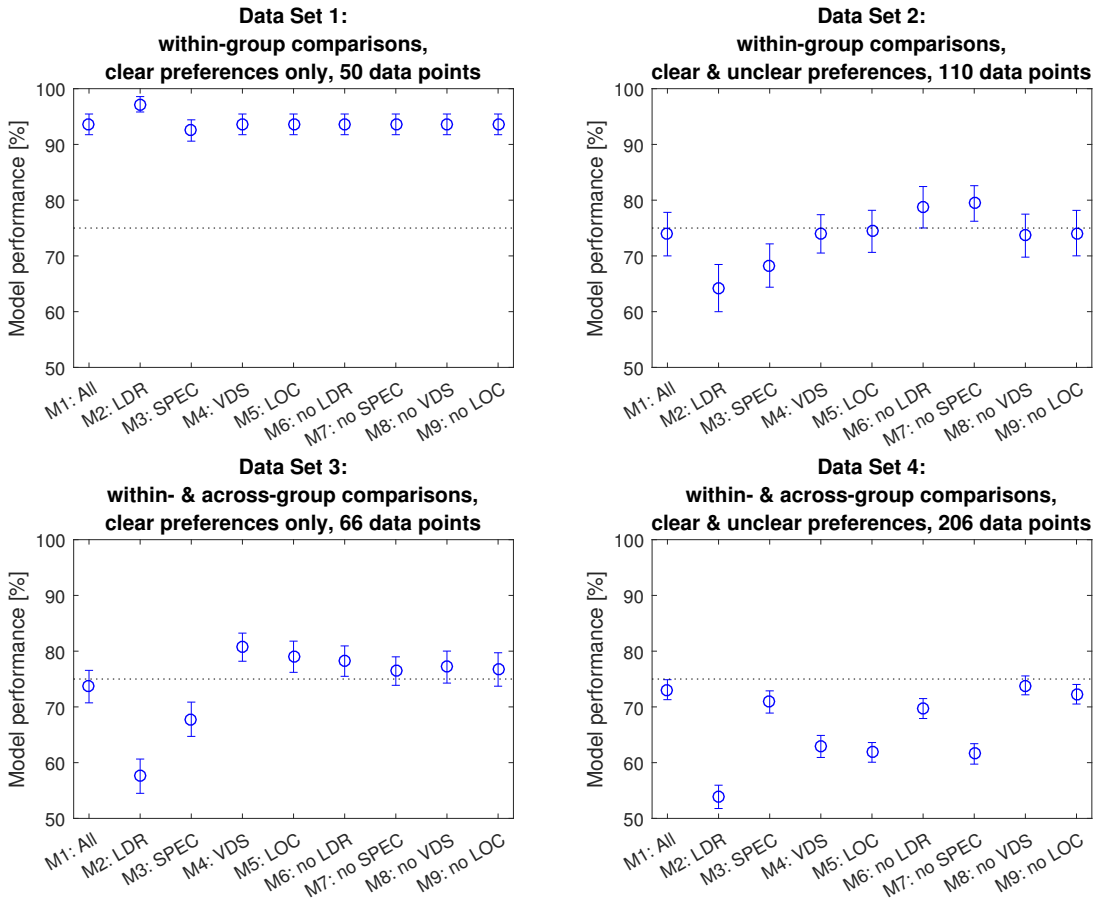


Figure 4.2: Performance of the nine models M1 to M9 for the four different data sets in terms of percentage correct classifications. The errorbars show the mean and 95% confidence intervals across 100 repetitions. As can be seen from the graph, all models perform above the chance level of 50%. The 75% lines indicate a rough differentiation into better and worse performing models.

of the different feature types, this threshold should be set relative to the overall model performance, which is actually calling for a soft decision margin rather than a strict threshold. Based on the results in Figure 4.2, we chose a decision margin as follows: If the performance of a model is in the order of 60% or below, than it is interpreted as bad performance; if the performance is in the order of 70% or above, then it is interpreted as good performance; if the performance is in-between, then its interpretation as good or bad performance depends on how the current model performance relates to the performance of the other models.

Figures 4.3 to 4.6 visualize the importance of each feature type for predicting the individual

comparison groups (Compression, EQ, Reverb, Positioning, Vocals, Mixed). The left panel of each figure shows the performance when the models are trained and validated on Data Set 2 (i.e. only within-group comparisons are included in the training); the right panel shows the performance when the models are trained and validated on Data Set 4 (i.e. within- and across-group comparisons are included). Each panel plots the performance of the full model with all feature types included (M1), the performance of that model in which the considered feature type is removed (M6, M7, M8, M9), and the performance of that model that uses only the considered feature type (M2, M3, M4, M5).

First, we discuss the performance of Model M1, as this is the same per Data Set in all four figures. The hypothesis for this model is that it performs well for all comparison groups. This hypothesis holds as the M1 performance does not drop below the 50% chance level for any group. However, in Data Set 4, the performance of M1 is quite low for groups Reverb and Positioning, as it is getting close to the chance level. For Data Set 2 this is not the case though. Furthermore, the performance of M1 for Compression is – unexpectedly – much worse for Data Set 2 than for Data Set 4. Hence the question of whether across-group comparisons are part of the training (Data Set 4) or not (Data Set 2) appears to have an impact on model performance, motivating that both cases are indeed reported.

Now, we discuss the LDR features (Figure 4.3), focussing first on the comparison group *Compression*. The expectation for Model M6 (no LDR) – bad performance – does not hold, as the performance is unexpectedly close to the full model (M1) performance for Data Set 4 and even higher than M1 in Data Set 2. Similarly surprising is the performance of M2 (LDR only) for the *Compression* group, as this was expected to work well, but shows a rather low performance (in the area of the decision margin) for both data sets. Focussing now on the comparison groups EQ, Reverb, and Positioning, the expectation for M6 – good performance – holds for all three groups only for Data Set 2. For Data Set 4, however, this holds only for the EQ group; for Reverb and Positioning M6 essentially fails, given a performance around chance level.

Thus, the benefit of the LDR features is highly ambiguous: On the one hand, these features alone are not very successful in characterizing the preference for Compression. On the other hand, these features appear to contribute to the characterization of the preference for Reverb and Positioning – at least for Data Set 4 – as they are raising the model performance from M6 to M1 above chance level.

Now, we discuss the SPEC features (Figure 4.4), focussing first on the comparison group *EQ*. The expectation for Model M7 (no SPEC) – bad performance for the *EQ* group – holds only for Data Set 4, surprisingly not for Data Set 2. The expectation for Model M3 (SPEC only) – good performance for the *EQ* group – holds for both data sets. Focussing now on the comparison groups Compression, Reverb, and Positioning, the expectation for M7 – good performance – holds for all three groups for Data Set 2. For Data Set 4, however, this holds

only for Compression; for Reverb and Positioning M7 essentially fails, given a performance around chance level, which is also lower than M1 performance.

Thus, the benefit of the SPEC features is proven as these features alone are successful in characterizing the preference for Equalization. In addition, these features also contribute to the characterization of the preference for Reverb and Positioning – at least for Data Set 4 – as they are raising the model performance from M7 to M1 above (Reverb) or further away (Positioning) from chance level.

Now, we discuss the VDS features (Figure 4.5), focussing first on the comparison group *Reverb*. The expectation for Model M8 (no VDS) – bad performance – holds only for Data Set 2, but not for Data Set 4, as the performance here is surprisingly good. Similarly, the expectation for Model M4 (only VDS) – good performance – holds only for Data Set 2, but again not for Data Set 4, as the performance here is in the order of the decision margin. Focussing now on the comparison groups Compression, EQ, and Positioning, the expectation for M8 – good performance – holds for all three groups for both data sets. However, for Data Set 4, it should be noted that for Positioning, M8 performs not only reasonably well, but also clearly better than M1.

Thus, the benefit of the VDS features is strongly dependent on the data set. For Data Set 2, the results support a benefit, since the VDS features alone are successful in characterizing the preference for Reverb, even though removing these features from the full model has only a slightly negative impact. For Data Set 4, however, the picture appears to be turned around: the VDS features alone are not successful, and removing these features from the full model has a moderate but clear positive impact.

Finally, we discuss the LOC features (Figure 4.6), focussing first on the comparison group *Positioning*. The expectation for Model M9 (no LOC) – bad performance – holds for Data Set 2, but not for Data Set 4. Similarly, the expectation for Model M5 (LOC only) – good performance – holds only for Data Set 2, but again not for Data Set 4. Focussing now on the comparison groups Compression, EQ, and Reverb, the expectation for M9 – good performance – holds for all three groups for Data Set 2. For Data Set 4, however, this holds only for Compression and EQ; for Reverb M9 shows rather low performance (in the area of the decision margin).

Thus, the importance of the LOC features is dependent on the data set, though in a slightly different way than the VDS features. For Data Set 2 the results support a benefit of the LOC features, as these features alone are successful in characterizing the preference for Positioning, even though removing these features from the full model has only a slightly negative impact. For Data Set 4, however, these features alone are not very successful, as the performance is in the area of the decision margin, while removing these features from the full model has a slightly negative impact, suggesting some small benefit.

4 Final quality model

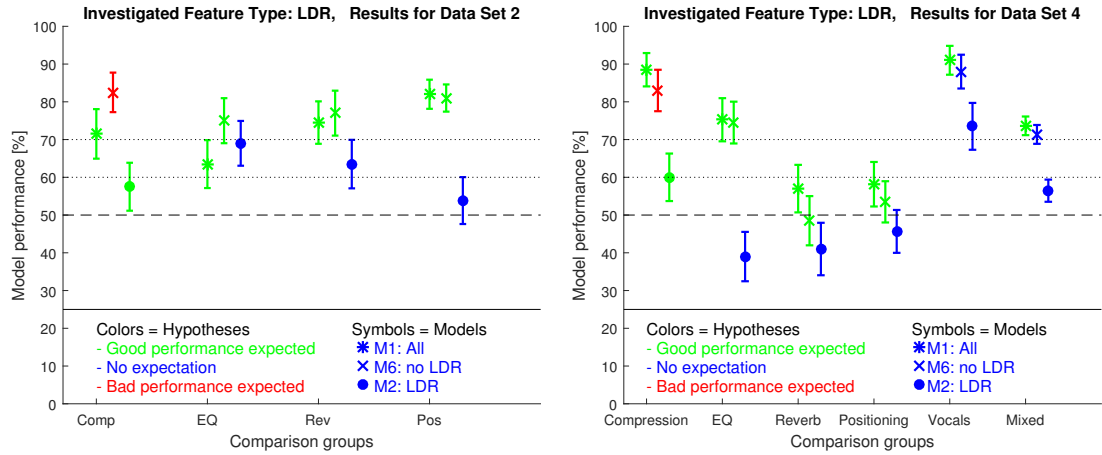


Figure 4.3: Modeling performance for the individual comparison groups: focussing of the models relevant for the impact of the LDR features (M1, M2, M6). The colors code the expectations according to Table 4.7. Left panel: results for Data Set 2 (within-group comparisons only). Right panel: results for Data Set 4 (within- & across-group comparisons).

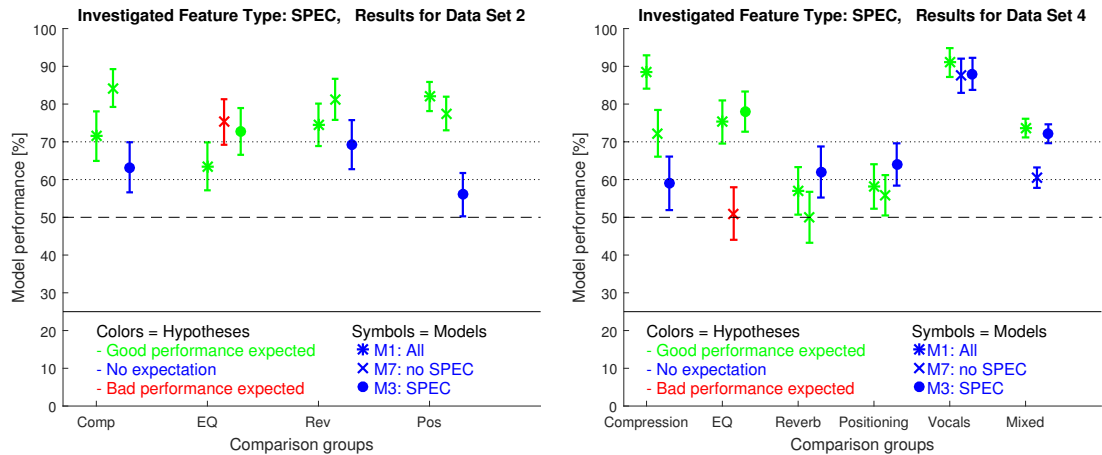


Figure 4.4: Modeling performance for the individual comparison groups: focussing of the models relevant for the impact of the SPEC features (M1, M3, M7). The colors code the expectations according to Table 4.7. Left panel: results for Data Set 2 (within-group comparisons only). Right panel: results for Data Set 4 (within- & across-group comparisons).

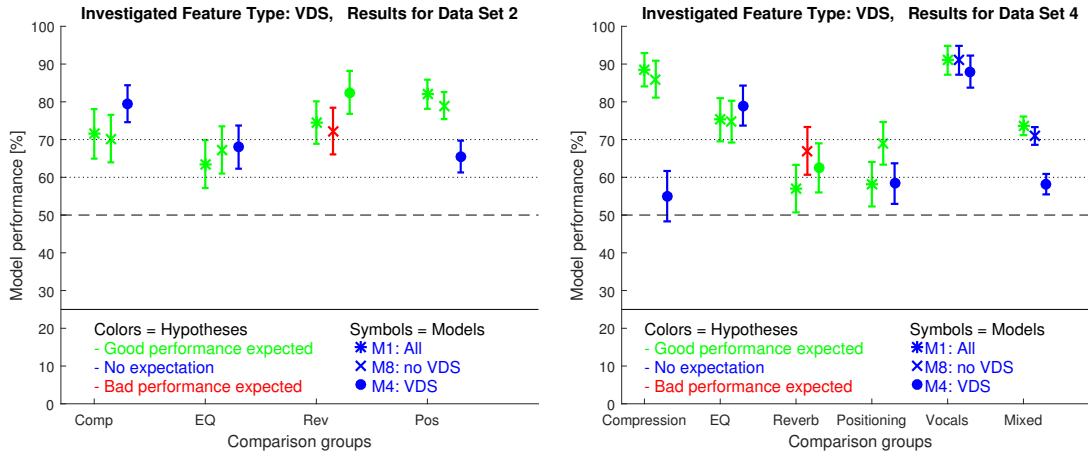


Figure 4.5: Modeling performance for the individual comparison groups: focussing of the models relevant for the impact of the VDS features (M1, M4, M8). The colors code the expectations according to Table 4.7. Left panel: results for Data Set 2 (within-group comparisons only). Right panel: results for Data Set 4 (within- & across-group comparisons).

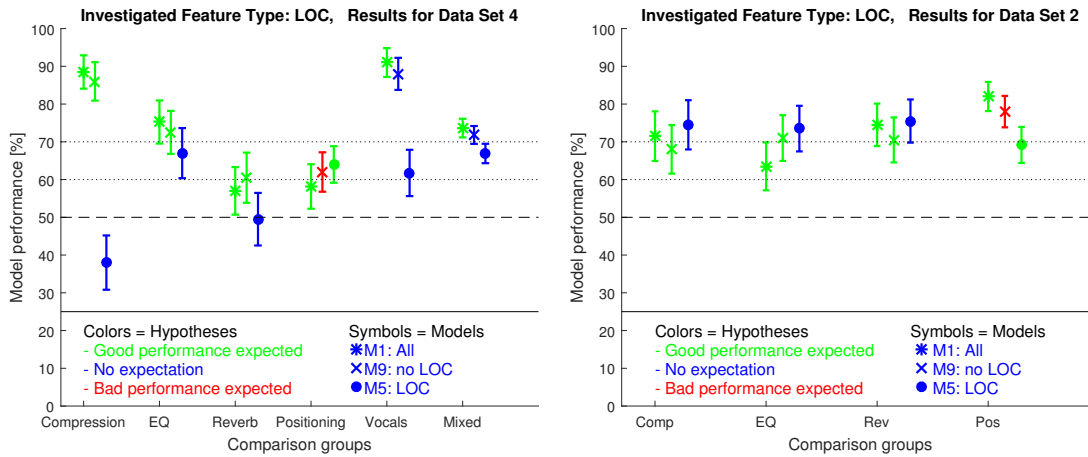


Figure 4.6: Modeling performance for the individual comparison groups: focussing of the models relevant for the impact of the LOC features (M1, M5, M9). The colors code the expectations according to Table 4.7. Left panel: results for Data Set 2 (within-group comparisons only). Right panel: results for Data Set 4 (within- & across-group comparisons).

4.2.8 Discussion

First, this study shows that pairwise preference ratings can principally be predicted, since the performance of the best models are clearly above the chance level of 50%. This proves the feasibility of the presented approach, but as values rarely reach high performance levels above 80 or even 90%, more work is necessary to obtain a robust model.

However, the foremost question is which performance can be actually achieved, given that the preference probability in the perceptual – that is the ground-truth – data is also low for the majority of cases. This portion of unclear preferences is 68% comparing Data Set 3 with 66 clear preferences vs. Data Set 4 with 206 data points in total. A rough indication for the present data set can be given by computing the average preference probability, which is just 0.62 for Data Set 4 and 0.59 for Data Set 4. While these numbers should not be misunderstood as the target percentages the model should achieve, such low numbers indicate that the task at hand is also very difficult for humans. This is inline with the general difficulty of assessing high-quality spatial audio systems.

Second, the observed dependency on the data set clearly shows that a robust model definitely needs more data. On the one hand, the number of music pieces should be increased; on the other hand, one might also consider to obtain ratings for more possible combinations of the mixed category (in which stimuli were compared that differed in two mixing parameters). The mixed comparison group appeared – as expected – to lead to a more challenging data set.

While such a conclusion in the sense of “more data is better data” is definitely reasonable here, the results also give an indication about the cleanness of the ground-truth data. Here, one can discuss which training data set is more appropriate: a data set with clean, i.e. prototypical examples per class, in order to best train the structure behind the problem, or a data set including also less prototypical examples, in order to obtain a model that is more robust against noisy data. While in times of big data analysis the latter approach is nowadays the typical choice, the present study suggests that prototypical data is favorable, at least for a small data set as in the present case: The per comparison group results are in most cases worse for the data set which includes also the mixed comparisons as training material than for the data set without those comparisons. This is especially visible for the results of the reverb features using the Van Dorp Schuitman model. Here, training with mixed comparisons led to the unexpected effect that including those features even decreased modeling performance, while training with non-mixed comparisons showed an added value of those features.

Third, this study confirmed our hypothesis to use a combination of different feature types, as different aspects of the mixing techniques differentiated the stimuli. While the LDR features appear to be expandable, the other three feature types SPEC, VDS and LOC were

beneficial in terms of modeling performance, even though they appeared to mutually cover the same aspects explaining the preference ratings, since one type could often fully be replaced by the other two.

To conclude, two directions of future work beyond addressing the issue of the limited data set is a) to find better performing features characterizing the preference ratings for compression and b) to investigate the dependency of the benefit of the VDS features and the issue of including the mixed group into the training data set.

5 Conclusions

In summary, the following conclusions can be drawn for the QoE-related work during year 3:

1. **Comparing different spatial audio reproduction systems:** The dedicated object-based mixing of audio for different reproduction methods showed to be effective and led to considerable quality differences between different reproduction methods. For modelling, it was decided to conduct additional sound quality – or since addressed in terms of preference – *Quality of Experience* tests with per-object variations in the mixing settings. The different variations lead to different preference ratings, forming the basis for developing a dedicated predictor. Since the amount of data that can be produced with listening tests in the lab is limited, we followed a feature-based approach rather than a machine-based training approach. The resulting model developed for preference prediction using the TWO!EARS framework proofs its principal suitability for *Quality of Experience* evaluation, as described in detail in Chapter 4.
2. **Finding the sweet spot for 5.1 surround:** Due to the difficulty to interpret the results, the work stream on sweet-spot identification was not considered in the more detailed modelling. The general paradigm and question of sweet spot identification can be considered as a highly relevant contributions to the community. To the knowledge of the authors, the “sweet spot” has not yet been defined so far in a technically and perceptually solid manner. Future tests following the same paradigm can be considered as a fruitful approach to approach a meaningful definition.
3. **Combined PC and MDS approach for listening preference and Scene-specific quality assessment:** It is obvious that, based on the now fully available TWO!EARS software framework, this workstream will be a very promising one to take up by other research projects, with TWO!EARS laying the foundation for a broader research approach on sound quality and *Quality of Experience* evaluation. As first proof of concept of a feature-based approach, the individual features localization and colouration were successfully modelled using the TWO!EARS framework, as described in D6.2.2. Hence, the rich set of knowledge sources made available with the TWO!EARS software – and possibly newly aggregated or added ones – can be considered to be suitable also when these and further quality features shall be used

in an integration to overall quality. Here, the in-depth software engineering and documentation across WP6 and the provided stimuli can serve as a starting point. In TWO!EARS, a dedicated MDS- and PC-based paradigm was applied in a pilot test described in this document, and ultimately not used for modelling beyond the existing per-feature model proofs described in D6.2.2, to focus on further work streams of *Quality of Experience*-related proof-of-concept. Using data from sensory evaluation campaigns (for example by labs where such data is already available) is considered as very promising, when the approaches undertaken within TWO!EARS as well as the rich set of modules from the TWO!EARS framework are used in combination. It is noted that the model type chosen in TWO!EARS for the proof of individual feature prediction was “no reference” for localisation, and “full reference” for coloration. For a no-reference coloration or general sound quality model based on features, a large amount of data will need to be collected to learn source-type-specific feature value-patterns. Considering the set of internal references that will need to be matched, the task of a generally valid sound quality model for such high-end systems as they have been addressed in TWO!EARS will require a set of follow-up research activities solely dedicated to such work.

The availability of data is a major issue when a variety of modelling efforts are to be addressed. Since no existing, openly available test data could be found that addresses the systems and evaluation paradigms targeted in TWO!EARS, the consortium had to produce all data itself. To this aim, a large number of assessment tests were conducted. As expected due to the fundamental research character related with topics such as the combination of the right mixing approach for a given reproduction system to achieve valid results, the modelling was focused on sub-parts of the collected data.

Relevant parts of the test data obtained in TWO!EARS, also of tests that were ultimately not used for modelling, has been made available open source, following the reproducible research paradigm of TWO!EARS. With the dedication to Open Science with a rich set of elements that is provided as “best practice information” to the community, the TWO!EARS group hopes to inspire and help other labs to apply a similar approach. If such an Open Science approach will generally be used in the areas of audio signal processing and sound quality evaluation, the amount of publicly available algorithms and test data will be increased significantly, creating ways of direct comparisons. With the open availability of a variety of features that can be used to implement quality evaluation algorithms, systematic quality model development and evaluation can be considered as a clear breakthrough. Here, the Auditory Modelling Toolbox has been a prominent example of an open source psychoacoustics model collection Søndergaard *et al.* (2011), Søndergaard and Majdak (2013). This has now been complemented by the more model-implementation-oriented TWO!EARS framework.

Bibliography

- Benoit, A., Callet, P. L., Campisi, P., and Cousseau, R. (2008), “Quality assessment of stereoscopic images,” in *IEEE International Conference Image Processing (ICIP)*, pp. 1231–1234. (Cited on page 5)
- Bitzer, J., LeBouf, J., and Simmer, U. (2008), “Evaluating perception of salient frequencies: Do mixing engineers hear the same thing?” in *124th Conv. Audio Eng. Soc.*, p. Paper 7462. (Cited on page 8)
- Blauert, J. and Raake, A. (2015), “Can current room-acoustics indices specify the quality of aural experience in concert halls?” *Psychomusicology: Music, Mind, and Brain* **25**(3), pp. 253–255. (Cited on page 57)
- Bradley, R. and Terry, M. (1952), “Rank analysis of incomplete block designs: I. The method of paired comparisons,” *Biometrika* **39**(3/4), pp. 324–345, URL <http://www.jstor.org/stable/10.2307/2334029>. (Cited on page 17)
- Brüggen, M. (2001a), “Coloration and binaural decoloration in natural environments,” *Acta Acustica united with Acustica* **87**, pp. 400–406. (Cited on page 29)
- Brüggen, M. (2001b), *Klangverfärbung durch Rückwürfe und ihre auditive und instrumentelle Kompensation*, dissertation.de, www.dissertation.de, D–Berlin. (Cited on page 29)
- Burstein, H. (1988), “Approximation Formulas for Error Risk and Sample Size in ABX Testing,” *J. Audio Eng. Soc.* **36**(11), pp. 879–883, URL <http://www.aes.org/e-lib/browse.cfm?elib=5124>. (Cited on pages 33 and 40)
- Chen, Y. W. and Lin, C. J. (2006), “Combining SVMs with various feature selection strategies,” *J. Agr. Biol. Envir. St.* . (Cited on page 63)
- Choisel, S. and Wickelmaier, F. (2007), “Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference,” *J. Acoust. Soc. Am.* **121**(1), pp. 388. (Cited on pages 9 and 17)
- David, E. E., Guttman, N., and van Bergeijk, W. A. (1959), “Binaural Interaction of High-Frequency Complex Stimuli,” *The Journal of the Acoustical Society of America*

- 31**(6), pp. 774–782. (Cited on pages 45, 46, and 52)
- De Man, B. and Reiss, J. D. (**2013**), “A pairwise and multiple stimuli approach to perceptual evaluation of microphone types,” *Aes134*, pp. Paper 8837. (Cited on page 37)
- Efron, B. (**1992**), “Bootstrap methods: another look at the jackknife,” in *Breakthroughs in Statistics*, Springer, pp. 569–593. (Cited on page 65)
- Francart, T., van Wieringen, A., and Wouters, J. (**2008**), “APEX 3: a multi-purpose test platform for auditory psychophysical experiments.” *Journal of neuroscience methods* **172**(2), pp. 283–293. (Cited on page 45)
- Fung, C. V. (**1996**), “Musicians’ and Nonmusicians’ Preferences for World Musics: Relation to Musical Characteristics and Familiarity,” *J. Res. Music Educ.* **44**(1), pp. 60–83. (Cited on page 9)
- Geier, M., Ahrens, J., and Spors, S. (**2008**), “The SoundScape Renderer : A Unified Spatial Rendering Methods,” *124th Conv. Audio Eng. Soc.* . (Cited on page 15)
- Giannoulis, D., Massberg, M., and Reiss, J. D. (**2013**), “Parameter Automation in a Dynamic Range Compressor,” *J. Audio Eng. Soc.* **61**(10), pp. 716–726. (Cited on page 9)
- Hold, C., , Nagel, L., Raake, A., and Wierstorf, H. (**2016a**), “Variations of pop mixes for Wave Field Synthesis,” URL <http://dx.doi.org/10.5281/zenodo.61000>. (Cited on page 15)
- Hold, C. and Wierstorf, H. (**2016a**), “Object-based audio scene files for variations of the spatial arrangement in pop mixes for Wave Field Synthesis,” URL <http://dx.doi.org/10.5281/zenodo.61110>. (Cited on page 15)
- Hold, C. and Wierstorf, H. (**2016b**), “Signal feeds for creating the music mixes for comparison of wave field synthesis, surround, and stereo,” URL <https://doi.org/10.5281/zenodo.55718>. (Cited on page 15)
- Hold, C., Wierstorf, H., and Raake, A. (**2016b**), “The Difference Between Stereophony and Wave Field Synthesis in the Context of Popular Music,” in *140th Conv. Audio Eng. Soc.*, p. 8. (Cited on page 15)
- Horbach, U., Karamustafaoglu, A., Pellegrini, R., Mackensen, P., and Theile, G. (**1999**), “Design and Applications of a Data-based Auralization System for Surround Sound,” in *106th Conv. Audio Eng. Soc.*, p. Paper 4976. (Cited on page 15)
- International Society for Music Information Retrieval (**2016**), “Overview of conference proceedings since 2000,” URL <http://www.ismir.net/conferences.html>, retrieved on November 26, 2016. (Cited on page 63)

- ITU-R BS.1534 (2015), *ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems*, International Telecommunications Union. (Cited on page 4)
- Jillings, N., De Man, B., Moffat, D., Reiss, J. D., and Stables, R. (2016), “Web Audio Evaluation Tool : A framework for subjective assessment of audio,” *2nd Web Audio Conference* . (Cited on page 32)
- Kendall, M. G. and Smith, B. B. (1947), “On the method of paired comparisons.” *Biometrika* **34**(Pt 3-4), pp. 324–345. (Cited on page 17)
- Kim, C. and Kohlrausch, A. (2016), “Investigations into Effects of Basilar Membrane Nonlinearity on Interaural Cues Using the Novel Modular Software Framework of the Two!Ears Project,” in *Association for Research in Otolaryngology 2016 MidWinter Meeting Abstract Book*, San Diego, p. 118. (Cited on page 44)
- Lebreton, P., Raake, A., Barkowsky, M., and Callet, P. L. (2013), “Perceptual preference of S3D over 2D for HDTV in dependence of video quality and depth,” in *IVMSP Workshop: 3D Image/Video Technologies and Applications*, Seoul, Korea. (Cited on page 5)
- Legendre, P. (2005), “Species Associations: The Kendall Coefficient of Concordance Revisited,” *J. Agr. Biol. Envir. St.* **10**(2). (Cited on page 19)
- Lepa, S., Ungeheuer, E., Maempel, H.-J., and Weinzierl, S. (2013), “When the medium is the message: An experimental exploration of medium effects on the emotional expressivity of music dating from different forms of spatialization,” in *8th Conference of the Media Psychology Division of Deutsche Gesellschaft für Psychologie (DGPs)*. (Cited on page 4)
- Mansbridge, S., Finn, S., and Reiss, J. (2012), “An Autonomous System for Multitrack Stereo Pan Positioning,” in *133rd Conv. Audio Eng. Soc.*, p. Paper 8763. (Cited on page 9)
- Mattila, V.-V. and Zacharov, N. (2001), “Generalized listener selection (GLS) procedure,” *In: Proc. 110th Audio Engineering Society (AES) Convention May 12-15, NL-Amsterdam*. (Cited on page 30)
- Misra, H., Ikbal, S., Sivadas, S., and Boulard, H. (2005), “Multi-resolution spectral entropy feature for robust ASR,” in *IN PROCEEDINGS OF IEEE INTERNATIONAL CONFERENCE ON ACOUSTIC, SPEECH, AND SIGNAL PROCESSING*, Citeseer. (Cited on page 64)
- Nagel, L. (2016), “Predicting preference in productions of popular music with an auditory model,” . (Cited on page 64)
- Pal, M. and Foody, G. M. (2010), “Feature selection for classification of hyperspectral data

- by SVM,” *IEEE Transactions on Geoscience and Remote Sensing* **48**. (Cited on page 63)
- Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011), “The timbre toolbox: Extracting audio descriptors from musical signals,” *The Journal of the Acoustical Society of America* **130**(5), pp. 2902–2916. (Cited on page 64)
- Pestana, P. and Reiss, J. D. (2014), “Intelligent Audio Production Strategies Informed by Best Practices,” in *53rd Int. Conf. Audio Eng. Soc.*, pp. S2–2. (Cited on page 9)
- Raake, A. (2016), “Views on Sound Quality,” in *Proceedings 22nd International Congress on Acoustics (ICA)*. (Cited on page 4)
- Raake, A. and Blauert, J. (2013), “Comprehensive modeling of the formation process of sound-quality,” in *Proc. IEEE QoMEX*, Klagenfurt, Austria. (Cited on page 7)
- Raake, A., Wierstorf, H., and Blauert, J. (2014), “A case for TWO!EARS in audio quality assessment,” in *Forum Acusticum*. (Cited on pages 6, 7, 30, 31, 32, 33, 35, and 36)
- Rumsey, F., Zielinski, S., Jackson, P., Dewhirst, M., Conetta, R., George, S., Bech, S., and Meares, D. (2008), “QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener,” in *125th Conv. Audio Eng. Soc.* (Cited on page 4)
- Rumsey, F., Zieliński, S., Kassier, R., and Bech, S. (2005), “On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality,” *Journal of the Acoustical Society of America* **118**(2), pp. 968–976. (Cited on page 4)
- Schoeffler, M. and Herre, J. (2016), “The relationship between basic audio quality and overall listening experience,” *The Journal of the Acoustical Society of America* **140**(3), pp. 2101–2112. (Cited on page 4)
- Schultze, A. (2016), “Der Sweet Spot in 5.0 Wiedergabesystemen in Abhängigkeit des Aufnahmeverfahrens und des visuellen Eindrucks des Zuhörers,” Ph.D. thesis, Technische Universität Berlin. (Cited on page 23)
- Skovenborg, E. (2014), “Measures of Microdynamics.” in *137th Conv. Audio Eng. Soc.* (Cited on pages 63 and 66)
- Skowronek, J. (2016), “Quality of Experience of Multiparty Conferencing and Telemeeting Systems – Methods and Models for Assessment and Prediction,” PhD thesis draft, accepted for defense. (Cited on page 67)
- Søndergaard, P. and Majdak, P. (2013), “The auditory-modeling toolbox,” in *The technology of binaural listening*, edited by J. Blauert, Springer, Berlin–Heidelberg–New York NY, chap. 2. (Cited on page 80)

- Søndergaard, P. L., Culling, J. F., Dau, T., Le Goff, N., Jepsen, M. L., Majdak, P., and Wierstorf, H. (2011), “Towards a binaural modelling toolbox,” in *Proceedings of Forum Acusticum*. (Cited on page 80)
- Spors, S., Wierstorf, H., Raake, A., Melchior, F., Frank, M., and Zotter, F. (2013), “Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State,” *Proceedings of the IEEE* **101**(9), pp. 1920–1938. (Cited on page 3)
- van Dorp Schuitman, J. (2011), “Auditory modeling for assessing room acoustics,” Ph.D. thesis, Delft University of Technology. (Cited on pages iv, 36, 53, 54, 55, and 56)
- van Dorp Schuitman, J., de Vries, D., and Lindau, A. (2013), “Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model,” *The Journal of the Acoustical Society of America* **133**(March), pp. 1572–1585, URL <http://www.ncbi.nlm.nih.gov/pubmed/23464027>. (Cited on pages 8, 36, 53, 56, and 64)
- Wältermann, M., Raake, A., and Möller, S. (2010), “Quality dimensions of narrowband and wideband speech transmission,” *Acta Acustica united with Acustica* **96**(6), pp. 1090–1103. (Cited on page 30)
- Wickelmaier, F. and Schmid, C. (2004), “A Matlab function to estimate choice model parameters from paired-comparison data,” *Behavior Research Methods, Instruments, & Computers* **36**(1), pp. 29–40, URL <http://dx.doi.org/10.3758/BF03195547>. (Cited on pages 17 and 33)
- Wierstorf, H. (2016), “Binaural room scanning files for a 56-channel circular loudspeaker array,” URL <http://dx.doi.org/10.5281/zenodo.55572>. (Cited on page 15)
- Wierstorf, H., Geier, M., and Spors, S. (2011), “A Free Database of Head Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances,” in *130th Conv. Audio Eng. Soc.*, London, UK, p. eBrief 6, URL <http://www.aes.org/e-lib/browse.cfm?elib=16564>. (Cited on page 15)
- Wierstorf, H., Hohnerlein, C., Spors, S., and Raake, A. (2014), “Coloration in Wave Field Synthesis,” in *AESC55*, pp. Paper 5–3. (Cited on page 8)
- Wierstorf, H., Raake, A., Geier, M., and Spors, S. (2013), “Perception of focused sources in wave field synthesis,” *Journal of the Audio Engineering Society* **61**(1/2), pp. 5–16. (Cited on page 30)
- Wilson, A. and Fazenda, B. (2016a), “Relationship Between Hedonic Preference and Audio Quality in Tests of Music Production Quality,” in *Proceedings IEEE 8th International Conference Quality of Multimedia Experience (QoMEX)*, pp. 1–6. (Cited on page 4)
- Wilson, A. and Fazenda, B. (2016b), “Relationship Between Hedonic Preference and Audio

- Quality in Tests of Music Production Quality,” in *QoMEX*, p. 6. (Cited on page 13)
- Wilson, A. and Fazenda, B. M. (2015), “101 Mixes: A statistical analysis of mix-variation in a dataset of multitrack music mixes,” in *139th Conv. Audio Eng. Soc.*, p. Paper 9398. (Cited on page 9)
- Wittek, H. (2015), “ORF Surround sound techniques, 2002,” URL <http://www.hauptmikrofon.de/stereo-3d/orf-surround-techniques>. (Cited on page 22)
- Zacharov, N. and Lorho, G. (2006), “What are the requirements of a listening panel for evaluating spatial audio quality?” in *Proc. Int. Workshop on Spatial Audio and Sensory Evaluation Techniques*. (Cited on page 30)
- Zacharov, N., Pike, C., Melchior, F., and Worch, T. (2016), “Next generation audio system assesement using the multiple stimulus ideal profile method,” in *QoMEX*. (Cited on page 5)