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

Deliverable 6.2.1

Quality of Experience test method specification

WP6 *

November 30, 2014

* 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 6.2.1
Document title:	QoE test method specification
Version:	1.0

Delivery date:	30. November 2014
Actual publication date:	30. November 2014
Dissemination level:	Restricted
Nature:	Other

Editor(s)/lead beneficiary:	Alexander Raake
Author(s):	Alexander Raake, Hagen Wierstorf, Fiete Winter, Sascha Spors, Chungeun Kim, Armin Kohlrausch, Thomas Walther, Jens Blauert, Tobias May, Patrick Danès
Reviewer(s):	Bruno Gas

Contents

1 Executive summary **1**

2 Specification of test method **3**
2.1 Introduction . . . 3
2.2 Modelling quality in Two!Ears . . . 5
2.2.1 Modelling *sound quality* . . . 5
2.2.2 Modelling *Quality of Experience* . . . 7
2.2.3 Specific modelling capabilities addressed in Two!Ears . . . 8
2.3 Analytic test methods: assessment of perceptual features . . . 9
2.3.1 Attribute elicitation . . . 10
2.3.2 Attribute ratings . . . 11
2.3.3 Methods to be used in Two!Ears . . . 12
2.4 Utilitarian test methods: assessment of *sound quality* . . . 13
2.4.1 Paired comparison test . . . 13
2.4.2 Modified MUSHRA . . . 15
2.4.3 Methods to be used in Two!Ears . . . 15
2.5 Assessment of *Quality of Experience* and indirect methods . . . 16
2.5.1 Methods to be used in Two!Ears . . . 17

3 Planned listening tests **19**
3.1 Investigation of different surround sound recording techniques . . . 19
3.2 Investigation of different spatial audio reproduction systems . . . 20

4 Contribution to Database in D1.1 **21**
4.1 Spatial fidelity . . . 21
4.2 Timbral fidelity . . . 23
4.3 Head movements . . . 24
4.4 Quality judgements for a music scene . . . 25

5 Conclusions **27**

Bibliography **29**

1 Executive summary

The developments in TWO!EARS require evaluation of the individual parts as well as of the entire software and test-bed functionality. This will be achieved based on two proof-of-concept applications that will be handled in work package 6 of TWO!EARS. Those are *Dynamic Auditory-Scene Analysis* with a special consideration of search-and-rescue-type scenarios, and *Quality of Experience* assessment.

For *Quality of Experience* assessment, the objective is first to develop suitable assessment methods for *sound quality* and *Quality of Experience* tests in an audio reproduction context. In a second step, these methods are applied to respective quality-related labelling of scenes presented via different spatial audio systems based on the collected quality test results. In later stages, the integrated TWO!EARS modules will be applied to *sound quality* and *Quality of Experience* assessment (for definitions see Sec. 2.1), covering multichannel loudspeaker systems driven with Wave Field Synthesis (WFS) or Near-Field Compensated Higher Order Ambisonics (NFC-HOA), as well as stereophonic systems. Specific objectives are the development of top-down model adaptation based on listeners' attention and active exploration, using both full-reference and no-reference modelling approaches. The considered evaluation criteria are perceptual attributes, *sound quality*, *Quality of Experience* and, if feasible, the criteria immersion, emotional response and task performance. For *Quality of Experience* and the afore mentioned criteria, the impact due to the system is assessed in terms of the meaning of the reproduced scenes and the role it plays for how the content is being perceived.

This deliverable mainly consists of two parts. First we define appropriate test methods and listening experiments, that deal with the perception of sound presented via different spatial audio systems. The second part is a description of all data already collected and contributed to the databases specified in D 1.1. Those data are stimuli and corresponding human labels from quality listening tests for different spatial audio systems, and in terms of specific quality-evaluation criteria.

This deliverable enables us to apply the specified test methods in various quality tests for spatial audio systems that will be carried out in the next two years of TWO!EARS.

2 Specification of test method

2.1 Introduction

This document targets the specification of *sound quality* and *Quality of Experience* tests carried out for data collection for the training and evaluation of the Two!Ears model. It covers *Quality of Experience* assessment as an application of the Two!Ears model as described in task 6.2. Before a meaningful specification of the test method can be made, the actual target measure of the test has to be defined.

When audio signals are played back via loudspeakers or headphones, different elements of the end-to-end chain from creation to presentation may impact what is being perceived by listeners (Spors *et al.*, 2013). Hence, for the design of audio technology it is of primary relevance to know how listeners will experience the produced acoustic scenes, that is, know the resulting *Quality of Experience*. *Quality of Experience* has been defined as (see Raake and Egger (2014)):

> ***Quality of Experience*** is the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state.

Like all perceptual events, quality 'happens' in the brain of the listener (Jekosch, 2005). In this context, two aspects of quality perception can be considered: When "perceived quality" directly addresses the acoustic scene in terms of the technical system, this manifestation has been coined as *quality (based on experiencing)* (Raake and Egger, 2014). Here, the person is aware of the technical system and assesses it directly, for example when a person evaluates different audio systems in a store that she considers to purchase, or when she is a test participant in an audio quality listening test. In the context of audio systems we will refer to this as ***sound quality***.

It the case of *Quality of Experience* it is noted that the listener does not necessarily need to be aware of how the underlying technology influences the listening experience. Obviously, real *Quality of Experience* according to this definition is hard to assess. In practice, research labeled as *Quality of Experience*-research is rather concerned with assessing *sound*

quality.

Mausfeld (2003) has highlighted the "dual nature" of perception, for example when watching a picture, or being in a virtual environment. Schoenenberg (2015) has underlined the validity of this view when assessing *Quality of Experience* in the context of mediated speech communication applications. A listener may focus (a) on the medium that is employed, in terms of an artifact, for example paying attention to the sound features related with a certain audio system when reproducing a musical piece, or (b) on the auditory scene presented to her. In case of everyday usage of audio technology, it can happen that degradations due to the technical system are attributed to the audio scene or a communication partner (Schoenenberg *et al.*, 2014). Those often cannot be measured in a test asking for *sound quality*, but are an important part of *Quality of Experience*. More details on the quality formation process in the mind of a person as it is assumed in this work can be found in Raake and Egger (2014).

In order to assess *sound quality* in a listening experiment different approaches are possible. The question we answer for Two!Ears in this deliverable is what listening test methods we should use to assess spatial audio systems. For an overview of methods see Bech and Zacharov (2006). The decision for particular methods is not only motivated by the listening tests themselves, but also by the ability to model perceived quality afterwards. Here, the test method must deliver valid and reliable results, but within a limited amount of time: The main objective of this task of Two!Ears is *Quality of Experience* model development, during the temporal boundaries of the project.

The test methods can be divided into *direct* and *indirect* ones. In *direct* methods listeners are directly asked to judge the perceived *sound quality* or *Quality of Experience* of a presented stimulus or to rate attributes that characterize the perceived scene or impact due to the technical system. Methods that assess other constructs related with perceived *sound quality* or *Quality of Experience* are referred to as *indirect* methods. These other constructs could be the *immersion* of a listener or the performance on specific tasks like *speech intelligibility*.

The direct approaches to *sound quality* assessment can further be divided into two classes. One is that of *analytic* methods for assessing the perceptual attributes that contribute to *sound quality* or *Quality of Experience*. Such attributes are typically analysed with respect to underlying perceptual dimensions. In the case of *sound quality* of spatial audio systems, these are mainly related to spatial and timbral aspects of the perceived signal. The other class comprises the *utilitarian* methods that ask directly for perceived quality, or for a ranking of the presented stimuli according to their quality, and can be used for mapping of underlying dimensions to preference (preference mapping, cf. Bech and Zacharov (2006)).

In the next sections we discuss the different approaches, and which of them we will use for

sound quality and *Quality of Experience* experiments.

2.2 Modelling quality in Two!Ears

In the introduction we have defined *sound quality* and *Quality of Experience* and how they can be assessed. Before we discuss the different test methods in more detail, we will take a look at quality assessment from a modelling perspective.

The goal of the TWO!EARS quality model is to tackle both *sound quality* of a given spatial audio system and *Quality of Experience* of a listener. For *sound quality*, the listener is instructed that the quality of the audio system is under investigation, and to directly rate quality or quality related attributes. For *Quality of Experience* the listener is only told to rate the quality of her experience, or to use content-related attributes to characterize the listening experience, and she is not indicated that the audio system is under test.

Sound quality and *Quality of Experience* models can be considered from two perspectives: If the goal is to study the perceptual processes during quality formation and reflect these in a perceptual model, this type of assessment can be considered as *subject-related* (Raake, 2006). In case that the replication of observed stimulus-response patterns and their relation with underlying stimulus-properties and/or properties of the employed audio technology are the primary goals, the respective assessment can be considered as *object-related*. In TWO!EARS, the main emphasis is on *subject-related* assessment, developing *sound quality* and *Quality of Experience* models functionally reflecting human quality formation. This is important to note, since this introduces some flexibility with regard to the underlying listening tests: Particular properties of the employed audio systems and their relation to quality are not the primary focus. This aspect will become even more apparent based on the considerations given in the following.

2.2.1 Modelling *sound quality*

During the assessment of the quality of a spatial audio system, the listener carries out a comparison of the perceived audio scene to a reference (Jekosch, 2005). This reference could be explicitly given, as it is the case for example in a classical MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) paradigm (ITU–R BS.1534-1, 2003). In the context of spatial audio systems, in most cases it is not possible to provide such an explicit reference: There mostly is no "real" sound scene that is intended to be recreated; for example, the goal in musical production is not to exactly recreate the sound field related with a given real, reference configuration of a band attended to in live-listening context.

Instead, the audio scene the listener experiences is artistically created with a special audio reproduction system in mind. Even if a live concert should be reproduced, the listener is not able to directly compare it to the real-life listening experience. When no explicit reference is provided to judge against, listeners use their internal reference they have built up during prior listening.

On top of this, the production process introduces an additional factor in the investigation of quality for different spatial audio systems. *Sound quality* is not only dependent on the audio system, but also on the artistic creation of the scene and the limits and possibilities of the artistic tools available for the different systems (Raake and Egger, 2014). To provide an example let us assume that we want to compare the two audio presentation techniques of two-channel stereophony and WFS. For stereophony lots of tools exist to create audio scenes, as well as experts that are trained to work with them. WFS not only provides the possibility of using the spatial domain more freely than possible with stereophony, but also introduces a paradigm-shift for audio production. At the current state, production for WFS requires an object-based approach for the creation and mixing process, which, in theory, has advantages over the channel-based panning approaches of stereophony. In reality, there is a lack of tools to create convincing audio scenes, and only a few experts are able to do this for WFS. If we want to compare both techniques in the context of the presentation of music, it is very likely that not only the technical systems, but also the production process of the music scene has a big influence on the judged *sound quality*.[1]

In order to exclude the influence of the production stage, different strategies are possible. In the context of sound field synthesis, the production process can be excluded by using only very simple scenes like a single point source. With such simple scenarios, the investigation is limited to simple perceptual attributes, too. With this approach, we investigated the localization and coloration for different sound field synthesis methods, summarized in Wierstorf (2014). However, in the context of more real-life listening conditions, it is of interest to use more complex scenes. This will not only be more realistic, but is also a requirement for highlighting differences between spatial audio systems, as the presentation of a single source is far from showing all possibilities or limitations of such systems.

Another strategy to avoid the explicit consideration of the production process is to use a spatial recording technique that can directly be reproduced by a spatial audio system. Here, we will focus on one reproduction setup, namely 5.0 stereophony and will investigate the usage of different microphone setups for the recording of the same scene. In this context, the perspective of a listening person can more easily be taken, since there is a larger variety of recording-setup/reproduction combinations. Hence, for the training of a perceptual

1 The same holds for the system-impact on *Quality of Experience*.

quality model, we can investigate single attributes of the perceived scene, the perceived *sound quality* or listening-session *Quality of Experience*, as well as indirectly assess the system-impact on *Quality of Experience* or immersion. The data can also easily be made available to the TWO!EARS model. The reproduction setup used in the experiments will be represented as a set of Binaural Room Impulse Responses (BRIRs) measured with the same Knowles Electronics Manikin for Acoustic Research (KEMAR) dummy head that was already used to measure Head-Related Impulse Responses (HRIRs) and BRIRs collected in D 1.1. The dummy head can rotate its head, and this principally allows head movements for scene exploration to be included into the modelling.

As mentioned at the beginning of this section, depending on the listening situation, listeners can compare the perceived scene to an explicit reference or to a set of internal ones. In the modelling, this translates into two different classes of models. One has access to the reference signal and is termed *full-reference* models, the other has no access to the reference signal and is termed *no-reference* models. The second modelling approach is more challenging than the first one. The first one can calculate difference features for both the scenes under test and the explicit reference, and respective differences between the auditory representations of the two scenes. A *no-reference* model first has to learn an internal reference based on auditory representations, which can then be used for difference calculation or direct prediction. In order to learn such an internal reference, the model needs to be implemented as a first running version. The work on the *full-reference* model started in the first year with the creation of a component that is able to predict coloration (Ende, 2014) (compare *Quality of Experience* Scenario 4 in D 6.1.1), whereas the work on the *no-reference* model started with the prediction of localization (*Quality of Experience* Scenarios 1–3 in D 6.1.1), that does not require a given reference signal, as it can be predicted as an absolute value with the help of learned features from measured HRIRs.

2.2.2 Modelling *Quality of Experience*

For the modelling of *Quality of Experience*, two different approaches exist. The first one is based on a running *sound quality* model, and the *Quality of Experience* is predicted by a combination of the *sound quality* modelling stage and an additional top-down component (realized as an expert in the software framework, compare D 3.2) that includes other factors. Those factors can be the liking/disliking of a given piece of music, the focus of attention of the listener, or the information whether the listener has listened to the presented audio system or audio piece before.

As the goal of this modelling stage is more to demonstrate that other factors beside *sound quality* can be important for *Quality of Experience* and can be reflected in a respective model, the main focus of the modelling is not on the ability to assess *Quality of Experience*

for arbitrary scenes like a human listener could do; the prior knowledge of a human person cannot realistically be built up during TWO!EARS. Instead, the experts for the liking/disliking of music for example will be purely based on results from a limited set of listening tests, and the respective experts likely operate in a mixed rule-based/statistics-based manner.

2.2.3 Specific modelling capabilities addressed in Two!Ears

Currently existing quality models have different limitations (see Raake and Blauert (2013) for more details): Explicit references are typically used, and internal references are considered only partially, as incase of the models by Beerends *et al.* (2013), Härmä *et al.* (2014), Jeong-Hun Seo and Choi (2013). The scene-based approach taken by some few models is rather rudimentary: foreground/background differentiation is used by Rumsey *et al.* (2008), some implicit aspects are reflected by the PEAQ model from Thiede *et al.* (2000) and in the model by Härmä *et al.* (2014). Active exploration is currently unavailable in models, and the peripheral models currently used may be improved using an extended set of features, and specifically be enhanced by top-down feedback (Blauert *et al.*, 2013).

We attempt to address these aspects with the TWO!EARS quality model. The general modelling paradigm has been outlined in Raake and Blauert (2013). It is based on the TWO!EARS model for interactive listening and auditory-scene analysis. Based on its modular architecture, the model is sought to functionally replicate human hearing and relevant aspects of cognition. The targeted sound-quality model shall introduce novel properties towards:

- Learned internal references rather than explicit reference signals, in principle enabling a no-reference sound quality model, or functionally adequate reference-adaptation for the case of full-reference model implementations.
- Scene-based quality assessment: Identification of scene and source types and respective adjustment of low-level processing as well as adjustment of the selected internal reference in the light of the given evaluation task.
- Explicit implementation of attentional processes based on the scene- and object-oriented paradigm.
- Integration with visual information, for example in terms of specific features of the scene.
- Active exploration.
  - ... targeting a specific analysis of certain low-level features exploited during

interactive quality evaluation (for example, based on behavioral patterns of listeners that evaluate acoustic scenes).

– . . . enabling the exploration of the scene, for example to identify the sweet-spot of a given sound reproduction system in a perceptual way. This is complementary to the experimental work described, for example, in Kim *et al.* (2013).

At first, *sound quality* evaluation will be addressed in terms of the definition given above. By importing knowledge from running research into immersion, emotional expression and listening experience (e.g. Lepa *et al.* (2013), Michael Schoeffler (2013)), attempts will be made to extend the paradigm from sound quality to actual *Quality of Experience* modelling.

After this overview of the different aspects we are trying to model and investigate in Two!Ears, the quality listening test methods to be applied in Two!Ears are now discussed in more detail.

2.3 Analytic test methods: assessment of perceptual features

The key aspect in *sound quality* evaluation of spatial audio systems is the fact that humans perceive acoustic scenes based on an auditory scene analysis. Making reference to learned world-knowledge, the listener associates the aural character of individual objects and/or the scene with internal references. It is noteworthy that these references may correspond to fixed schemata, as in the case of the telephone or stereo reproduction: Here, prior listening experience has lead to internal references of their own kind, see Jekosch (2005). Now, when the *sound quality* of a reproduction system is assessed, a comparison of perceived features with regard to desired features is performed.

Analytic test methods are able to identify and investigate those perceived features. This is especially important in the context of model development: Quality models that assess *sound quality* by predicting single features and combining them seem to be a promising approach, see, for example, Wältermann (2013) in the context of speech quality.

The systems addressed in Two!Ears are multi-channel loudspeaker sound field synthesis, headphone-based binaural synthesis or stereophonic systems. In this case, quality is determined primarily by timbral and spatial features, and by spectro-temporal artifacts (Rumsey, 2002a, Berg and Rumsey, 2003, Rumsey *et al.*, 2005, Wittek, 2007, Wierstorf *et al.*, 2013a, Spors *et al.*, 2013). The simplest approach for analytic quality evaluation is the direct usage of scales that represent the underlying quality dimensions. Direct rating of attributes after explaining listeners the meaning of the scales using example stimuli has been applied to speech and audio quality evaluation Wältermann (2013).

For stereophonic sound reproduction, Rumsey (2002a) found quality to be determined by *timbral* and *spatial fidelity*, which explained 70% vs. 30% of the quality variance, respectively (Rumsey *et al.*, 2005). Extensive spatial fidelity assessment has been conducted, for example in Berg and Rumsey (2003), Rumsey *et al.* (2005), Wittek (2007), Wierstorf (2014). The timbre related with different WFS systems has been studied in Wierstorf *et al.* (2014). Artifacts may, for example, be introduced by spatial aliasing as it occurs in practically realized multichannel-audio presentation (Wierstorf *et al.*, 2013b). It should be noted, that added artifacts may lead to additional auditory streams processed separately from the underlying scene (Thiede *et al.*, 2000).

Preference ratings often show a high correlation with the attribute *loudness*. To avoid this correlation, stimuli are scaled with a listening experiment to have the same loudness before evaluation, as the overall loudness is not a fixed property of spatial audio systems and can easily be adjusted by listeners in real-life contexts. Hence, in Two!Ears, we will not investigate the implication of loudness on perceived quality. If the audio scene consists of different sources, it can happen that a spatial audio system changes their relative loudness. In this case, we will explicitly include loudness in the investigation, since per-source adaptation is possible only in case of an object-based audio-recording and -reproduction paradigm.

For complex perceptual features related with different spatial audio systems, where it is not obvious that they can be explained solely by one of the attributes *coloration*, *spaciousness*, and "affected by artifacts", the attributes will first be identified before the rating stage (*attribute elicitation*). The methods used in both cases are explained in the following subsections.

2.3.1 Attribute elicitation

When a dimension-based quality modelling strategy is applied, quality ratings are combined with attribute-ratings, as outlined above. If an experiment shows that the *sound quality* cannot be predicted alone by the chosen attributes it indicates that the set of used attributes is not complete, and missing perceptual dimensions have to be identified. There are different ways to do so. Nicol *et al.* (2014) discussed various methods and highlighted that the task is challenging for spatial audio, as the used stimuli in the elicitation test have to contain all possible degradations in order to register all underlying attributes/dimensions. The different elicitation methods can be grouped into two classes: 1) non-verbal elicitation, and 2) verbal elicitation. Note that *verbal* attribute elicitation pertains to the category of *qualitative* test methods, and scales produced using the attributes lead to *quantitative methods*. In turn, *non-verbal* attribute elicitation as described below directly enables quantitative multidimensional analysis. However, in this case, verbal attributes to describe the identified dimensions are not available.

Non-verbal elicitation & multidimensional analysis These test methods ask for a rating of differences between the presented stimuli. The most prominent method is Multi-Dimensional Scaling (MDS). The listeners are asked to rate the difference between two presented stimuli. Those difference ratings are afterwards used to calculate a space that is able to map all of the differences, and specifies the dimensions spanning that space. If the listeners give different weight to the different dimensions, this can be analysed by using the INDSCAL-method (Carroll and Chang, 1970). A classical example of applying MDS to audio systems is given by Gabrielsson and Sjögren (1979).

Another similar method is the Perceptual Structure Analysis (PSA) (Choisel and Wickelmaier, 2006). Here, a triad of stimuli is presented, and the listener is asked to rate whether the first two stimuli have a common feature that the third one does not have. At the end, similar to MDS, a space with a number of perceptual dimensions is created.

The advantage of these methods is that they work without the need of verbally describing the auditory perception, which can be very difficult (Mason *et al.*, 2001). The disadvantage is that the dimensions found in the experiments have to be labeled before they can be used in follow-up attribute rating experiments with a reduced set of attributes. One solution is to accompany the non-verbal elicitation experiment with a debriefing round in which listeners are asked to name possible attributes describing their listening experience during the elicitation (Gabrielsson and Sjögren, 1979, Wickelmaier and Ellermeier, 2007). Another option is to carry out both MDS-type and attribute-rating tests (Mattila, 2001, 2002, Wältermann, 2013, Bech and Zacharov, 2006).

Verbal elicitation Verbal elicitation can be performed on a single listener basis or directly using group discussions. The most prominent methods are Descriptive Analysis (DA) (Zacharov and Koivuniemi, 2001), Repertory Grid Technique (RGT) (Berg and Rumsey, 1999), Free-Choice Profiling (FCP) (Lorho, 2005), Individual Vocabulary Profiling (IVP) (Lokki *et al.*, 2011) or expert interviews (Lindau *et al.*, 2014). The disadvantages of these methods are that they require a verbal description of the perception, typically rely on expert listeners, and are mostly rather time consuming.

2.3.2 Attribute ratings

After relevant attributes have been found for perceptually assessing spatial audio systems, they are investigated by applying them in scaling experiments. Depending on their nature, different methods are possible for the assessment. Spatial attributes like *localization* and *locatedness* may directly be assessed via pointing methods (Wierstorf, 2014). Attributes that can be rated on a scale with two endpoints can be assessed via a MUSHRA based

approach (ITU–R BS.1534-1, 2003). This has the advantage over methods like the *semantic differential* that listeners are able to compare the different systems under test directly and are more sensitive to small differences. We did an assessment via a MUSHRA-like approach for coloration in WFS (Wierstorf *et al.*, 2014). In that case an explicit reference was given for the rating, but the method can also be modified to exclude a reference (Palacino *et al.*, 2012) and be used in the same way.

For perceptually complex attributes such as *apparent source width* we found that the assessment via the modified version of MUSHRA is too complicated for listeners (Busuru, 2013). In this case, listeners were only able to assess the differences between the stimuli via direct drawings of the perceived width (see also Hess (2006)).

2.3.3 Methods to be used in Two!Ears

Due to the existing research on attributes in the context of spatial audio quality (see the attributes discussed below) we will focus on the usage of non-verbal elicitation methods, mainly MDS. This pragmatic choice allows us to spend more time on the modelling of the results. Subsequent identification of verbal attributes will first employ post-test panel discussions with test participants. If these shall not lead to convincing labels, an explicit verbal attribute elicitation will be considered.

For the attribute ratings we will apply a MUSHRA-type approach with or without known reference. In addition, where a direct assessment is possible like for localization or the apparent source width, we might consider to use such direct assessment. As the work on surround sound systems (Rumsey, 2002b) and our own work on sound field synthesis have shown (Wierstorf *et al.*, 2013b), the attributes that contribute the most to *sound quality* can be grouped into three different classes: Spatial attributes, timbral attributes, and artifacts-related attributes. Depending on the spatial audio systems under investigation we will directly let listeners assess attributes from those categories.

Depending on the used stimuli we will follow two different approaches: If the dimensions to be scaled in the attribute rating are known, direct rating will be used, as Wältermann (2013) has done successfully for speech quality. If the stimuli are new and/or the underlying dimensions are not exactly known, the attributes first have to be elicited. In this case, we will start with the MDS together with post-test panel discussions in order to elicit attributes. Afterwards, those attributes will be used in the attribute rating experiments.

2.4 Utilitarian test methods: assessment of *sound quality*

Utilitarian test methods are used when the overall *sound quality* of the system is under investigation. Utilitarian tests normally only ask for a single value, rating the *sound quality* or the preference-order of the presented systems.

Often times, explicit reference stimuli are used in *sound quality* tests, for example in the tests typically conducted for audio coding quality evaluation. Known methods of this type are MUSHRA (ITU–R BS.1534-1, 2003) and BS.1116 (ITU–R BS.1116-1, 1997). Here, MUSHRA targets intermediate quality differences, and BS.1116 small differences between, for example, audio coding algorithms. For *sound quality* tests addressing a wide range of quality levels, single-stimulus methods such as the 5- or 9-point Absolute Category Rating (ACR) tests are typically used (ITU–R BS.1284-1, 2003, ITU–T Rec. P.800, 1996). Here, specific stimuli are often presented as hidden references that are not identified as such to the test participants. If no reference is available, the MUSHRA method can be modified by removing the explicit reference and using only hidden anchors (Palacino *et al.*, 2012). Modified MUSHRA has also been introduced for measuring single attributes, see Sec. 2.3. To test preference of presented stimuli, a Paired-Comparison (PC)-test is normally applied (Wickelmaier *et al.*, 2009).

The PC and modified MUSHRA methods are explained in more detail in the following sections, both methods will also be chosen in Two!Ears as is summarized in Sec. 2.4.3. The section on PC also discusses why it might be better to ask the listener for preference instead of direct quality ratings.

2.4.1 Paired comparison test

In a first listening test investigating different degradations that are typical of WFS we have tested the paired comparison test paradigm (Raake *et al.*, 2014). The test stimuli consisted of an audio scene that was composed of three different sources, two guitars and a singer. Even though the test in principle included a hidden reference stimulus, the audio scene presented via single loudspeaker, the results show that it was not rated to be the preferred stimulus. The outcome of the test showed that the listeners preferred the scenes in which the guitars were presented with WFS over the scenes where this was not the case. The results highlight that the no-reference paradigm is very important for assessing the quality of spatial audio systems.

Another concern with direct quality ratings is related with the criteria based on which quality is being evaluated: Research from the domain of image and video quality assessment indicates that single-stimulus tests targeting quality ratings may lead to unexpected results, due to specific quality-evaluation criteria used by the test participants, which are not

related with the perceptual advantage of stereoscopic 3D vision, but rather with the notion of "signal clarity", as explained in the following (Lebreton *et al.*, 2013): When stereoscopic 3D and 2D video sequences are assessed in the same quality test, test viewers often rate 2D sequences higher than 3D sequences, even if the coding bitrate and resolution imply high quality in both cases. This effect is assumed to be due to quality being rated in terms of "pictorial quality" or signal clarity, where 2D images or videos may appear to be superior for current 3D imaging technology. However, the actual advantage of 3D, namely to provide binocular depth information, is not considered by the test participants in this type of ratings. Similar observations were made for 2D-videos of different resolutions, either scaled or unscaled to the target (equal or higher) resolution of the test screen: the non-scaled sequences were typically rated highest – obviously again in terms of pictorial quality. In practice, however, down to certain lower-bound resolutions, users of video streaming services such as Youtube typically scale the video to full-screen, indicating the *preference* for this setting.

To circumvent these methodological problems, image and video quality research has re-adopted the approach of PC preference tests in recent years (e.g. Handley (2001), Benoit *et al.* (2008), ITU-T Contr. COM 12-C192-E (2011), Lebreton *et al.* (2013)). This way, other criteria than pictorial quality are considered in the more holistic evaluation as well, without however providing dedicated references. This may appear surprising at first, but becomes more clear when the actual task for the test-participants is re-considered: In ACR-tests or so-called Subjective Assessment Methodology for Video Quality (SAMVIQ)-tests (the video-equivalent to MUSHRA), the assessment task explicitly addresses the overall quality for a given condition, either rated in an "absolute" manner (ACR), or in comparison to the reference (SAMVIQ). The underlying quality-concept in the mind of the viewers is that of "goodness" or "excellence" in terms of, as stated above, pictorial quality, since this is the most apparent "quality" dimension. In contrast, paired comparison preference tests provide a more holistic evaluation, where the simple task lies in answering which of two stimuli is the preferred one. Respective results from 3D-image and -video and 2D-video quality research clearly prove this aspect, and for example 2D sequences are not at all generally "preferred" over 3D-sequences, see for example Lebreton *et al.* (2013).

Using methods such as the Thurstone-Mosteller or Bradley-Terry models enables the transformation of the PC-data to a continuous quality scale (e.g. Handley (2001)). It is obvious that full paired comparisons for a set of N stimuli with $N \cdot (N-1)$ comparisons, or if the stimulus order is excluded $N \cdot (N-1)/2$ comparisons, reduces the number of test stimuli that can be assessed with meaningful effort. Here, however, alternative square designs are available that lead to quite efficient tests (e.g. Li *et al.* (2012)).

Similar considerations apply to the case of spatial-sound quality assessment. Here, too, neither ACR-type tests nor tests with explicit references appear to be meaningful ap-

proaches. In a comparable way as for image and video tests, the choice of test conditions and specific focus on dedicated features may yield a bias of the results. While we have applied a full paired comparison in our first test (Raake *et al.*, 2014), we will employ more efficient test designs in future tests (Li *et al.*, 2012). However, a final decision on the larger-scale usage of the latter approach can only be taken after applying it to the task of spatial audio quality evaluation.

2.4.2 Modified MUSHRA

The full PC test introduced in the last subsection has one major drawback: if the number of different stimuli is high, the number of paired comparisons to be performed by the listener will increase drastically. If the number of stimuli is too large for a paired comparison test, and alternative designs such as square designs do not produce meaningful results, the modified MUSHRA, introduced in Sec. 2.3, may be a good alternative test method. Here, the listener is asked to rate the presented stimuli according to *sound quality* or their perceived preference. As several stimuli are presented in one set, this method is faster than a full paired comparison. Practical differences between such MUSHRA-type and non-full PC tests will be considered in the light of the finally chosen stimulus sets. Another advantage of the modified MUSHRA test is that it maps the ratings directly to a quality scale, whereby the results from different PC-tests are not always comparable on the same scale. This is a generally important point for the data collection in Two!Ears: The test design has to reflect the need for a unique quality scale, a research topic addressed by work package 6.2 in year two.

2.4.3 Methods to be used in Two!Ears

As already mentioned, in the context of spatial audio systems, an actual reference is not available in most cases. On the other hand, the test often includes stimuli that only show small differences. This has led us to the decision to use the following two methods for utilitarian tests in Two!Ears.

For utilitarian-type *sound quality* evaluation, PC tests and modified MUSHRA tests will be employed. Here, the priority will be on PC tests. Modified MUSHRA tests will be used complementing the PC-tests, in two ways:

- To create a general quality scale, as the target value to be estimated by the model. For stimuli that have large quality differences we may consider to use ACR in this context.
- To complement the PC-tests in case that too many conditions need to be tested for

an efficient use of PC-tests, and reduced, e.g. square designs show not to be usable.

2.5 Assessment of *Quality of Experience* and indirect methods

In Sec. 2.1 we discussed that both the *sound quality* of the underlying spatial audio system and the presented audio content (music piece) contribute to *Quality of Experience*. In order to classify their influences, three sets of listening experiments will be planned. All of them ask the listener to rate the preference for the presented stimuli. In one listening test the content of the presented stimuli will be fixed in order to force the listener to rate the *sound quality* of the underlying audio systems. These tests are the ones described in Sec. 2.4. In another experiment the audio system will be fixed and the content will be changed. Here, the listeners are asked to rate their liking of the different contents. In a third experiment, a mixture of content and audio systems will be presented to get a rating of the influence on *Quality of Experience* of mixing content and audio systems, compare Schoeffler *et al.* (2014). Those test will be planned, based on pre-tests in the second year and will be performed in the third year. Since their use for model data collection has to be validated in pre-tests at first, these tests are not listed in Chap. 3.

Besides *Quality of Experience*, a selected set of indirect assessment will be applied, if feasible. In the case of localization, the head movements made by the listener have already been used to judge the difficulty the listeners have with different conditions (Wierstorf *et al.*, 2012). In addition, the movements of the listener will be tracked in experiments where the listener is allowed to freely inspect the listening area for different spatial audio systems. This is of interest as the modelling approach allows the model to perform exploratory movements. For spatial audio systems and a listening situation where immersion is the goal, immersion-related ratings will be collected, and the relation with *sound quality* ratings will be investigated. Note that here, too, pre-tests for validation are needed, which will be considered for year two.

Another indirect method we will approach is the usage of a cognitively complex test task that requires auditory scene analysis, like the identification of the number of sound sources or the identity of single sources. For example, some of our previous research has employed the identification of human speakers for assessing a spatial audio presentation in the context of telephony applications (Raake *et al.*, 2007). This topic is currently discussed together with the partners, as such tasks are also relevant for the dynamic auditory scene analysis application of the model, where benchmarking against human performance is an important but difficult testing topic (Task 6.1). The idea is to run human listening experiments for such tasks while the respective audio scene is provided via different spatial audio systems. The question under investigation then is whether the audio system degrades the performance of the listener. In this regard, however, the investigated systems may be

too little different so as to assess performance-differences.

Under discussion at the moment is the inclusion of listening tests assessing the emotional response of listeners during their listening via different spatial audio systems. It was shown by different authors that the perceived emotions of the listener are dependent on the used spatial audio system (Västfjäll, 2003, Lepa *et al.*, 2014).

2.5.1 Methods to be used in Two!Ears

The direct assessment of *Quality of Experience* asking for overall listening experience (Schoeffler *et al.*, 2014) as described at the beginning of this section will be carried out accompanying *sound quality* tests. As the research conducted in work package 6 was shifted slightly to task 6.1 in the first year, we were not able to conduct pre-tests on the topic of indirect test methods so far. Which of the above criteria may be fruitful when assessing spatial audio systems can only be decided when respective test set-ups have been assessed in first informal tests. This task will be addressed early in year two.

3 Planned listening tests

The listening tests planned in the second year of Two!Ears target two different types of spatial audio systems. The first one will focus on different recording techniques for 5.0 surround sound, whereas the other will focus on the reproduction side by investigating the influence of different reproduction systems on a spatial audio scene.

The listening tests will include first preference ratings in a PC test, followed by MDS of the same stimuli. The PC test might be accompanied by a modified MUSHRA test as outlined in Sec. 2.4.3. For the cases where a sufficiently high number of different audio contents is available, direct *Quality of Experience* assessment as described in Sec. 2.5 will be considered, thus complementing the *sound quality* tests by overall listening experience assessment (Michael Schoeffler, 2013, Schoeffler *et al.*, 2014).

In addition, in pre-tests we will study different indirect assessment methods, for example targeting the impact of the different audio systems on the emotional impact of the audio scene.

3.1 Investigation of different surround sound recording techniques

As discussed in Sec. 2.2, one way to discard the influence of the artistic creation of a sound scene is to focus on recordings of real scenes using techniques that can directly be reproduced using a spatial audio system. To this aim, we are currently collecting the material of two large recording sessions that are freely available for research. One is a recording session performed during the International Conference on Spatial Audio 2011 in Detmold.[1] Here, seven different recording techniques were employed to simultaneously record nine different music performances. In this case not all of the recordings target surround sound as reproduction system, but also binaural or WFS reproduction. The second recording session was carried out during the surround sound seminar at ORF (Vienna) and solely targets surround sound reproduction.[2] Here, nine different recording techniques were used

1 `http://www.eti.hfm-detmold.de/musikproduktionen/aufnahmen_icsa` (in German only)
2 `http://www.hauptmikrofon.de/index.php?option=com_content&view=article&id=68`

in parallel to record two different music performances.

We plan to use the different recordings to run listening tests with a real 5.0 setup in a listening room and with binaural simulations of the same 5.0 setup. This is of interest especially for model development, since the model will only have the simulated binaural signals as an input.

3.2 Investigation of different spatial audio reproduction systems

In order to compare different spatial audio systems like mono reproduction, two-channel stereophony, 5.0 surround sound, WFS, and NFC-HOA, the presented audio scene cannot be extracted directly from a recording, but have to be adjusted by the experimenter. We will create a simple spatial audio scene, consisting of a few sources comparable to the three sources we have used in the quality test presented in (Raake *et al.*, 2014). The goal is to achieve a spatial complexity of the audio scene so that the different audio systems will have an influence on its spatial perception.

For WFS and NFC-HOA we will also compare different setups with a varying number of loudspeakers, as we did already for the assessment of the perceptual attributes *localization* and *coloration* (Wierstorf, 2014). The listening experiments will be performed with actual loudspeaker setups at TU Berlin and University of Rostock. Both groups have loudspeaker arrays to run WFS or, only at TU Berlin, NFC-HOA.

4 Contribution to Database in D1.1

A huge amount of tests were run at TU Berlin during the last years investigating spatial and timbral fidelity perception for the two sound field synthesis methods WFS and NFC-HOA. This work was and will further be complemented in Two!Ears, and the data is part of the public database that is described in D 1.1.

The following chapter outlines the different experiments that were performed, lists the contributed data, and discusses the connection to the final Two!Ears *sound quality* model.

4.1 Spatial fidelity

One of the key aspects of *sound quality* for spatial sound presentation techniques is the spatial fidelity provided by such systems. For a 5.1 stereophonic system, Rumsey *et al.* (2005) found that spatial fidelity explained around 30% of the variance of overall perceived quality.

During the last two years we have conducted a large number of listening experiments to investigate the spatial fidelity for different WFS and NFC-HOA systems. The experiments are summarized in Wierstorf (2014). The main focus of the investigation was on the localization and the locatedness of an auditory event synthesized by such systems.

It was found that the spatial fidelity was different for the different systems. For WFS and loudspeaker distances below 20 cm, no impairment of the spatial fidelity is perceived by the listeners, and the localization performance is the same across a large listening area. For larger inter-loudspeaker spacings, localization errors arise at all listening positions. For NFC-HOA the localization errors are much more listener-position-dependent and are worse for large loudspeaker spacings than in case of WFS. The results are summarized in Fig. 4.1.

The data are also of special interest for the development of the Two!Ears model, as the results for a focused source synthesized with WFS show: Here, there clearly is a need to incorporate a precedence effect model into the Two!Ears system. This development is

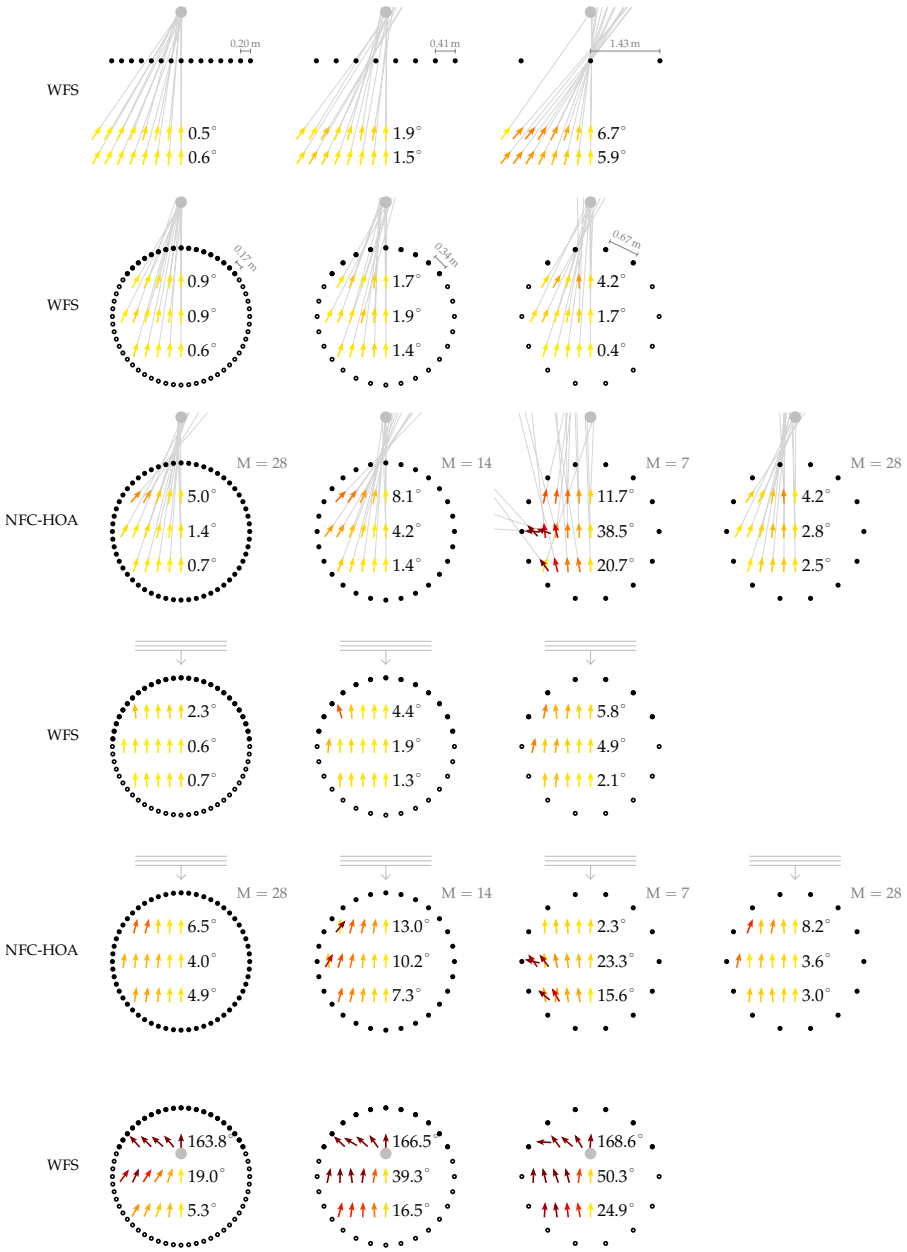

Figure 4.1: Average localization results for all experiments. The black symbols indicate loudspeakers, the grey ones the synthesized source. At every listening position, an arrow is pointing into the direction of the corresponding auditory event perceived by the listeners on average. The color of the arrow displays the absolute localization error, which is also summarized as an average beside the arrows for every row of positions. The average confidence interval for all localization results is 2.3°. M indicates the maximum order used for NFC-HOA.

supported by the definition of *Quality of Experience* Scenario 3 in D 6.1.1. Further, the auditory event starts to split into two for some of the NFC-HOA systems. At the moment, the localization stage of the Two!Ears model is not able to detect such a splitting. This effect should be included, and this task is specified by *Quality of Experience* Scenario 1 in D 6.1.1.

Another directly usable property of the performed listening tests lies in the ability to use head movements within the Two!Ears model. In the tests, head-movements were allowed. Due to the usage of dynamic binaural synthesis during the experiments and the possibility of the Two!Ears Binaural Simulator[1] developed in work package 1 to use exactly the same input BRIRs that were used during the experiments, the same head-motion patterns and resulting stimuli can be used in the model.

The data from the experiments described here are included into the database presented in D 1.1. The stimuli used in the experiments are listed under Database Entries #6 – #11. The human labels (test results) from the listening experiments, including the direction and locatedness of the auditory event are listed under the Database Entries #26 – #31.

4.2 Timbral fidelity

Another key contribution to overall *sound quality* is the color of the sound produced by spatial audio systems, in previous research addressed in terms of timbral fidelity (Rumsey *et al.*, 2005), here referred to as *coloration*.

We investigated the perception of coloration for a point source synthesized with different WFS systems. We varied the number of used loudspeakers in a large range, and investigated two different listening positions. Further, three different kinds of audio materials were used: pink noise pulses, speech, and music. The experiment is described in detail in Wierstorf *et al.* (2014), Wierstorf (2014).

The results for noise and speech are summarized in Fig. 4.2. It was found that all WFS systems that employ a loudspeaker distance larger than 2 cm suffer from coloration, and that this effect has to be covered by any model that shall be able to predict the *sound quality* for a WFS system.

The experimental data described here are included into the database presented in D 1.1. The stimuli used in the experiments are listed under Database Entry #12. The human labels (test results) from the listening experiments are listed under the Database Entry

1 `https://github.com/TWOEARS/binaural-simulator`

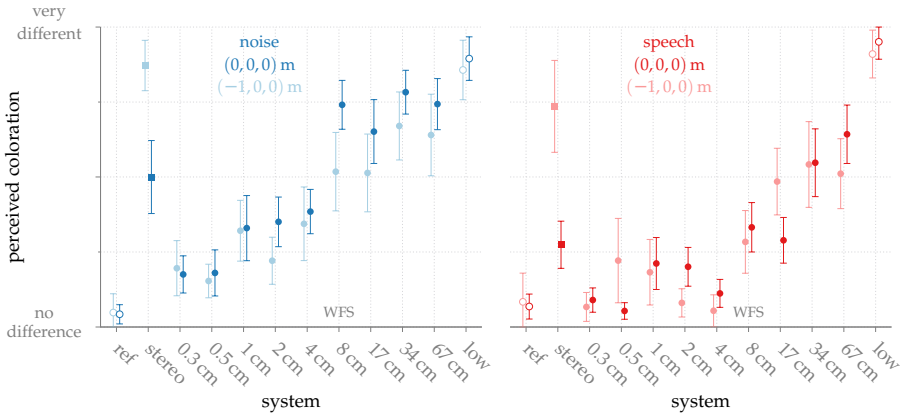

Figure 4.2: Average results with confidence intervals for perceived coloration. Dark colors show results for the central listening position, lighter colors for the off-center position.

#32. They will be used for the development of the coloration model as described by *Quality of Experience* Scenario 4 in D 6.1.1.

4.3 Head movements

One of the new aspects that is integrated into the Two!Ears model is the possibility to actively explore a given audio scene. One aspect here is the ability to rotate the head. In order to investigate different head rotation patterns during localization and compare them to the behavior of humans during the same task, listener head movements have been recorded during a localization test. The localization task was performed for real loudspeakers placed in a listening room and compared to localization of the same loudspeaker simulated with BRIRs measured at exactly the same position as the human listener was placed at, or with HRIRs from an anechoic recording. The experiment is further described in Wierstorf *et al.* (2012).

Figure 4.3 shows the results for three different loudspeaker positions and room conditions. It can be seen that the overall rating time[2] increases with the deviation of the simulation from the real setup, an increasing amount of head movements around the actual position was made.

The corresponding stimuli are presented in Database Entry #5 and #16 of D 1.1. The head movements of the single listener is part of Database Entry #24.

2 The rating was done by turning the head into the perceived direction and pressing a button, which marks the moment when the rating time was recorded.

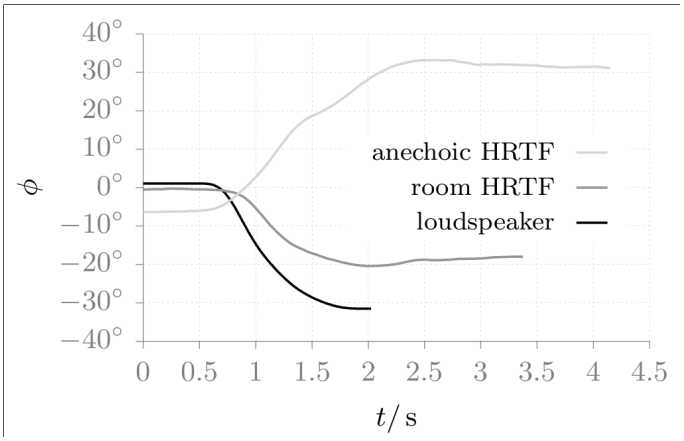

Figure 4.3: Head movements of a single listener between two source locations.

4.4 Quality judgements for a music scene

In another experiment already mentioned in Sec. 2.4, in a paired comparison test listeners rated which of two given audio stimuli was of higher quality. The two stimuli always consisted of the same audio scene with two guitars synthesized to the right and left of the listener, and a singer in front of the listener. Each part of this audio scene was degraded independently by synthesizing it as a point source or a focused source with different WFS setups. The degradations were introduced independently in order to investigate the influence of the single objects on the perceived overall quality. The experiment is further described in Dierkes (2014), Raake *et al.* (2014).

Figure 4.4 summarizes the results. The two main findings are that the reference condition, including no degradations (condition 8.), was not rated as the one having the best quality. The other result indicates that the degradation of the singer has a more unpleasant impact than the degradation of the guitars, highlighting the fact that overall quality depends on the degradation of individual sources.

The stimuli are presented as Database Entry #22, the quality ratings in Database Entry #33.

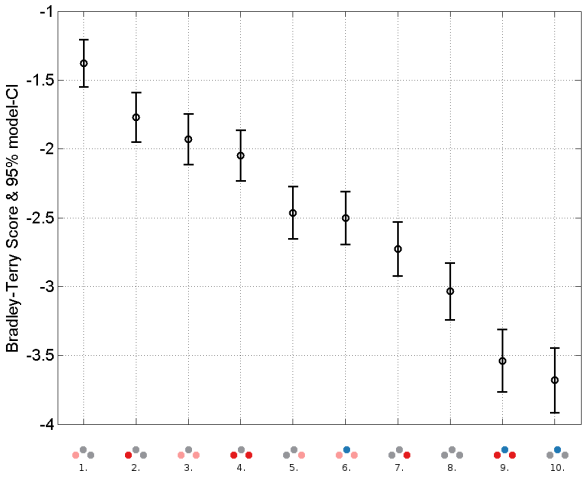

Figure 4.4: Conditions and the resulting ranking from the scene quality test. The paired comparison matrix has been transformed to a continuous Bradley-Terry score. High scores correspond to high quality and vice-versa. Red points highlight a modification of the corresponding guitar by the spatial audio system, blue points a modification of the singer and grey points no modification at all. This means that the reference condition was rated to only be the eighth best condition.

5 Conclusions

In this Deliverable, we have described the framework for data collection for developing the Two!Ears *Quality of Experience* Application and proof of concept. Based on a definition of *sound quality* and *Quality of Experience*, the targeted model development was briefly outlined, highlighting the specific research directions where Two!Ears shall make significant contributions. These considerations are applied to discuss different evaluation approaches for Two!Ears data collection. In this context, it was pointed out that tests with a hidden or explicit reference are not always suitable for spatial audio quality evaluation, since an actual reference mostly does not exist.

Two main evaluation criteria will be addressed in Two!Ears: (i) *sound quality* and (ii) *Quality of Experience*.

For *sound quality*, large parts of the existing literature and test data that is available to us is based on a "fidelity" paradigm, with regard to the underlying dimensions *color of sound / coloration*, *spaciousness* and *artefacts*. Hence, for model development, this approach will be considered for the full-reference type modelling strategy followed in Two!Ears. The different test results already available to the consortium stem from own tests and have been contributed to D 1.1, or stem from external parties known to individual Two!Ears partners. The latter data cannot be shared, and hence is considered to be used only for model development. Since the present document specifies new tests and past tests for which data has been contributed to the database in D 1.1, the aforementioned additional, external model training data is not further discussed here.

New test data in terms of a *utilitarian* test paradigm will be collected using Paired-Comparison (PC) tests, targeting preference ratings. The PC-tests will complemented by modified MUSHRA tests, which shall primarily serve to find a common, uni-dimensional quality scale used as target for model predictions. In both cases, no explicit references will be used. Ratings will be given by comparing two or multiple stimuli, depending on the applied test method (PC-tests: 2 stimuli; modified MUSHRA: sets of around 12 stimuli). It is expected that a dimension-based modelling approach will be advantageous over a paradigm where *sound quality* is directly predicted from the aural or scene-analysis-based cues available from work packages 2 and 3. Hence, *utilitarian* test methods will be complemented by *analytic* test methods. To this aim, Multi-Dimensional Scaling (MDS) will be used, combined with subsequent attribute-labelling, as further detailed in the main

part of this Deliverable.

For *Quality of Experience*, a combination of direct and indirect tests will be targeted. Here, it will be key to keep a good balance between a valid handling of the data collection, but without spending resources that must be reserved for the modelling activity. To reflect these boundary conditions, direct tests focusing on the overall liking of audio listening sessions with varying contents will be run complementing the *sound quality* tests. Here, instructions to test participants and test design will be such that the attention will not be on the underlying technical system, but focus on the overall audio listening experience. A third group of tests will target the judgment of overall listening experience, too, but combining different systems and different contents/scenes.

In addition, a selected set of meaningful indirect tests and data collections will be carried out. Here, data collection specifically addresses the recording of listener movements in reproduced sound fields as indirect indicator for either *sound quality* or *Quality of Experience*. In addition, specific aural features attended to may be identified based on the data. As QoE-related evaluation criteria, *immersion* and *emotional response* will be considered. The specification of these indirect *Quality of Experience* assessment methods will be addressed during year 2 based on pre-tests still to be carried out. Here, it must be investigated, whether available test methods are effective and efficient to collect data actually indicative of system-specific differences in *immersion* and *emotional response*. Only if respective data can be collected in a reduced amount of time, it will be possible to accommodate such tests in addition to the planned *sound quality* tests and the respective model development.

Here, it is noted that the pre-tests initially planned for year one for indirect *Quality of Experience* assessment have been shifted to year 2. These resources were shifted from Task 6.2 to Task 6.1 and other work packages. This shift reflects the more direct integration of work package 6.1-scenarios with the different Two!Ears model components, both for model development and systematic scenario-based evaluation. This represents a significantly improved approach over the initial project plans. Similarly, the object-oriented and modular implementation of the development system required an increased effort for work packages 1 and 2. Since the resulting scenario- and model-framework can be considered as more solid than initially planned, the resources missing in year 1 can be re-directed to work package 6.2 during years 2 and 3.

Bibliography

Bech, S. and Zacharov, N. (**2006**), *Perceptual Audio Evaluation - Theory, Method and Application*, John Wiley & Sons, Chichester, England. (Cited on pages 4 and 11)

Beerends, J. G., Sschmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., and Keyhl, M. (**2013**), "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech QualityMeasurement Part II – Perceptual Model," *J. Audio Eng. Soc.* **61**(6), pp. 385–402. (Cited on page 8)

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 14)

Berg, J. and Rumsey, F. (**1999**), "Spatial Attribute Identification and Scaling by Repertory Grid Technique and other methods," in *16th International Conference on Spatial Sound Reproduction, Audio Engineering Society*, pp. 51–66. (Cited on page 11)

Berg, J. and Rumsey, F. (**2003**), "Systematic Evaluation of Perceived Spatial Quality," in *Proc. 24th Conf. Audio Eng. Soc.* (Cited on pages 9 and 10)

Blauert, J., Kolossa, D., Obermayer, K., and Adiloglu, K. (**2013**), "Further challenges – and the road ahead," in *The technology of binaural listening*, edited by J. Blauert, Springer, chap. 18. (Cited on page 8)

Busuru, M. A. (**2013**), "Influencing the Auditory Source Width of Virtual Sources in Wave Field Synthesis," Bachelor thesis. (Cited on page 12)

Carroll, J. D. and Chang, J.-J. (**1970**), "Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition," *Psychomatrika* **35**(3), pp. 283–319. (Cited on page 11)

Choisel, S. and Wickelmaier, F. (**2006**), *"Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound,"* **54**(9), pp. 815–826. (Cited on page 11)

Dierkes, J. (**2014**), "Qualität räumlicher Audiowiedergabe: ist es szenenspezifisch oder objektspezifisch?" Bachelor thesis. (Cited on page 25)

Ende, C. (**2014**), "Auditorische Modellierung der Klangverfärbung in der Wellenfeldsynthese," Bachelor thesis. (Cited on page 7)

Gabrielsson, A. and Sjögren, H. (**1979**), "Perceived sound quality of sound-reproducing systems," *Journal of the Acoustical Society of America* **65**(4), pp. 1019–1033. (Cited on page 11)

Handley, J. C. (**2001**), "Comparative analysis of Bradley-Terry and Thurstone-Mosteller model of paired comparisons for image quality assessment," . (Cited on page 14)

Härmä, A., Park, M., and Kohlrausch, A. (**2014**), "Data-driven modeling of the spatial sound experience," in *Audio Engineering Society Convention 136*. (Cited on page 8)

Hess, W. (**2006**), "Time-Variant Binaural-Activity Characteristics as Indicator of Auditory Spatial Attributes," Ph.D. thesis, Ruhr-UniversitÃd't Bochum. (Cited on page 12)

ITU–R BS.1116-1 (**1997**), *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*, International Telecommunication Union, CH–Geneva. (Cited on page 13)

ITU–R BS.1284-1 (**2003**), *General methods for the subjective assessment of sound quality*, International Telecommunication Union, CH–Geneva. (Cited on page 13)

ITU–R BS.1534-1 (**2003**), *Method for the Subjective Assessment of Intermediate Quality level of coding systems*, International Telecommunication Union, CH–Geneva. (Cited on pages 5, 12, and 13)

ITU–T Rec. P.800 (**1996**), *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, CH–Geneva. (Cited on page 13)

ITU-T Contr. COM 12-C192-E (**2011**), *Comparison of theACR and PC evaluation methods concerning the effects ofvideo resolution and size on visual subjective ratings*, International Telecommunication Union, Geneva. (Cited on page 14)

Jekosch, U. (**2005**), *Voice and Speech Quality Perception — Assessment and Evaluation*, Springer, D–Berlin. (Cited on pages 3, 5, and 9)

Jeong-Hun Seo, K.-M. S., Sang B. Chon and Choi, I. (**2013**), "Perceptual Objective Quality Evaluation Method for High- Quality Multichannel Audio Codecs," *J. Audio Eng. Soc.* **61**(7/8), pp. 535–545. (Cited on page 8)

Kim, C., Mason, R., and Brookes, T. (**2013**), "Head Movements Made by Listeners in Experimental and Real-Life Listening Activities," *J. Audio Eng. Soc.* **61**(6), pp. 425–438. (Cited on page 9)

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*. (Cited on page 14)

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 9)

Lepa, S., Weinzierl, S., Maempel, H.-J., and Ungeheuer, E. (**2014**), "Emotional Impact of Different Forms of Spatialization in Everyday Mediatized Music Listening : Placebo or Technology Effects?" in *136th Convention of AES*. (Cited on page 17)

Li, J., Barkowsky, M., and Callet, P. L. (**2012**), "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment," in *ICIP*. (Cited on pages 14 and 15)

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S. (**2014**), "A spatial audio quality inventory for virtual acoustic environments (SAQI)," in *EEA Joint Symposium on Auralization and Ambisonics*, pp. 3–5. (Cited on page 11)

Lokki, T., Pätynen, J., Kuusinen, A., Vertanen, H., and Tervo, S. (**2011**), "Concert hall acoustics assessment with individually elicited attributes," *Journal of the Acoustical Society of America* **130**(2), pp. 835–849. (Cited on page 11)

Lorho, G. (**2005**), "Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction," in *119th Convention of the AES*, Paper 6629. (Cited on page 11)

Mason, R., Ford, N., Rumsey, F., and de Bruyn, B. (**2001**), *"Verbal and Nonverbal Elicitation Techniques in the Subjective Assessment of Spatial Sound Reproduction,"* **49**(5), pp. 366–384. (Cited on page 11)

Mattila, V.-V. (**2001**), *Perceptual Analysis of Speech Quality in Mobile Communications*, vol. 340, Doctoral Dissertation, Tampere University of Technology, FIN–Tampere. (Cited on page 11)

Mattila, V.-V. (**2002**), "Descriptive Analysis and Ideal Point Modelling of Speech Quality in Mobile Communications," *In: Proc. 113th Audio Engineering Society (AES) Convention* **October 5-8, USA–Los Angeles**. (Cited on page 11)

Mausfeld, R. (**2003**), "Conjoint representations and the mental capacity for multiple simultaneous perspectives," in *Looking into pictures: An interdisciplinary approach to pictorial space*, edited by H. Hecht, R. Schwartz, and M. Atherton, MIT Press, pp.

17–60. (Cited on page 4)

Michael Schoeffler, J. H. (**2013**), "About the Impact of Audio Quality on Overall Listening Experience," in *Proceedings of the Sound and Music Computing Conference (SMC)*, pp. 58–53. (Cited on pages 9 and 19)

Nicol, R., Gros, L., Colomes, C., Warusfel, O., Noisternig, M., Bahu, H., Katz, B. F. G., and Simon, L. S. R. (**2014**), "A roadmap for assessing the quality of experience of 3D audio binaural rendering," in *EAA Joint Symposium on Auralization and Ambisonics*, pp. 100–106. (Cited on page 10)

Palacino, J., Nicol, R., Emerit, M., and Gros, L. (**2012**), "Perceptual assessment of binaural decoding of first-order ambisonics," in *Acoustics*. (Cited on pages 12 and 13)

Raake, A. (**2006**), *Speech Quality of VoIP – Assessment and Prediction*, John Wiley & Sons Ltd, Chichester, West Sussex, UK. (Cited on page 5)

Raake, A. and Blauert, J. (**2013**), "Comprehensive modeling of the formation process of sound-quality," in *Proc. IEEE QoMEX*, Klagenfurt, Austria. (Cited on page 8)

Raake, A. and Egger, S. (**2014**), "Quality and Quality of Experience," in *Quality of Experience. Advanced Concepts, Applications and Methods*, edited by S. Möller and A. Raake, Springer, Berlin–Heidelberg–New York NY, chap. 2. (Cited on pages 3, 4, and 6)

Raake, A., Spors, S., Ahrens, J., and Ajmera, J. (**2007**), "Concept and Evaluation of a Downward-Compatible System for Spatial Teleconferencing using Automatic Speaker Clustering," *In: Proc. 10th Int. Conf. on Spoken Language Processing (Interspeech 2007 – ICSLP), BE-Antwerp* . (Cited on page 16)

Raake, A., Wierstorf, H., and Blauert, J. (**2014**), "A case for Two!Ears in audio quality assessment," in *Forum Acusticum*. (Cited on pages 13, 15, 20, and 25)

Rumsey, F. (**2002**a), "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *Journal of the Audio Engineering Society* **50**(9), pp. 651–666. (Cited on pages 9 and 10)

Rumsey, F. (**2002**b), *"Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm,"* **50**(9), pp. 651–66. (Cited on page 12)

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 Convention of the Audio Engineering Society*. (Cited on page 8)

Rumsey, F., Zieliński, S., Kassier, R., and Bech, S. (**2005**), "On the relative im-

portance 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 pages 9, 10, 21, and 23)

Schoeffler, M., Conrad, S., and Herre, J. (**2014**), "The Influence of the Single/Multi-Channel-System on the Overall Listening Experience," in *55th AES Conference on Spatial Auido*. (Cited on pages 16, 17, and 19)

Schoenenberg, K. (**2015**), "The Quality of Mediated Conversations under Transmission Delay," Ph.D. thesis, Technische Universität Berlin, to appear. (Cited on page 4)

Schoenenberg, K., Raake, A., and Koeppe, J. (**2014**), "Why are you so slow? - Misattribution of transmission delay to attributes of the conversation partner at the far-end," *International Journal of Human-Computer Studies* **72**, pp. 477–487. (Cited on page 4)

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 pages 3 and 9)

Thiede, T., Treurniet, W., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K., and Feiten, B. (**2000**), "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.* **48**, pp. 3–29. (Cited on pages 8 and 10)

Västfjäll, D. (**2003**), "The Subjective Sense of Presence, Emotion Recognition, Virtual Environments," *CyberPsychology & Behavior* **6**(2), pp. 181–188. (Cited on page 17)

Wältermann, M. (**2013**), *Dimension-based Quality Modeling of Transmitted Speech*, Springer, Berlin–Heidelberg. (Cited on pages 9, 11, and 12)

Wickelmaier, F. and Ellermeier, W. (**2007**), "Deriving auditory features from triadic comparisons," *Perception and Psychophysics* **69**(2), pp. 287–297. (Cited on page 11)

Wickelmaier, F., Umbach, N., Sering, K., and Choisel, S. (**2009**), "Comparing Three Methods for Sound Quality Evaluation with Respect to Speed and Accuracy," in *126th Convention of the AES*. (Cited on page 13)

Wierstorf, H. (**2014**), "Perceptual Assessment of Sound Field Synthesis," Ph.D. thesis, Technische Universität Berlin, to appear. (Cited on pages 6, 10, 11, 20, 21, and 23)

Wierstorf, H., Geier, M., Raake, A., and Spors, S. (**2013**a), "Perception of Focused Sources in Wave Field Synthesis," *J. Audio Engineering Soc.* **61**(1/2), pp. 5–16. (Cited on page 9)

Wierstorf, H., Geier, M., Raake, A., and Spors, S. (**2013**b), *"Perception of Focused Sources*

in Wave Field Synthesis," **61**(1), pp. 5–16. (Cited on pages 10 and 12)

Wierstorf, H., Hohnerlein, C., Spors, S., and Raake, A. (**2014**), "Coloration in Wave Field Synthesis," in *Audio Engineering Society Conference on Spatial Audio 55*. (Cited on pages 10, 12, and 23)

Wierstorf, H., Spors, S., and Raake, A. (**2012**), "Perception and evaluation of sound fields," in *59th Open Seminar on Acoustics*. (Cited on pages 16 and 24)

Wittek, H. (**2007**), "Perceptual differences between wavefield synthesis and stereophony," Ph.D. thesis, University of Surrey. (Cited on pages 9 and 10)

Zacharov, N. and Koivuniemi, K. (**2001**), "Audio Descriptive Analysis & Mapping of Spatial Sound Displays," in *Proceedings of the 2001 International Conference on Auditory Display*. (Cited on page 11)