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

Deliverable 6.2.2

QoE model software, first version

WP6 *

November 24, 2015

* 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 6.2.2
Document title:	QoE model software, first version
Version:	1

Delivery date:	30. November 2015
Actual publication date:	01. December 2015
Dissemination level:	Restricted
Nature:	Report

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):	Jonas Braasch, Dorothea Kolossa, Bruno Gas, Klaus Obermayer

Contents

1 Executive summary

The TWO!EARS model will be evaluated in the area of two possible applications. Those are *Dynamic Auditory Scene Analysis* and *Quality of Experience* (QoE). The first application is discussed in D 6.1.2, this document focusses on the work towards the QoE application of the model.

For the *Quality of Experience* assessment, the work is not only focussed on the actual model development, but also on acquiring data from listening tests. For this purpose we defined appropriate test methods in D 6.2.1 and ran different listening tests since the last deliverable. In the first chapter we will discuss the general way forward to a QoE model in TWO!EARS. At its current state we focus on the prediction of single attributes like coloration and localisation we investigated in different sound field synthesis methods like Wave Field Synthesis and Higher-Order Ambisonics. This will be done in a full-reference manner. We will run listening tests at the beginning of the final year in order to collect ground truth data for the final non-reference approach, where the reference can be learned and might be adapted using top-down feedback. Those tests are described as QoE-3 and QoE-4 in section 3.2 in Deliverable D 6.1.2.

All test data collected during year 2 is contributed to the public TWO!EARS database, the current state of which is described in Deliverable D 1.2.

2 Towards a model for QoE

2.1 Introduction

The Two!Ears project aims to develop an intelligent, active computational model of auditory perception and experience that operates in a multi-modal context. Evaluating *Quality of Experience* (QoE) is one of the two proof-of-concept applications of the model. In Two!Ears, QoE evaluation focuses on the listening to spatial audio systems.

The present report summarises the first quality model developments until the end of the second project year. It builds on the *Quality of Experience test method specification* provided in Deliverable D 6.2.1. Following the steps described in D 6.2.1, the model development targets sound quality (that is, quality based on experiencing) and Quality of Experience (Raake and Egger, 2014). During year 2, the primary focus of model development has been on sound quality evaluation. To this aim, the evaluation of coloration and preferred sound quality were investigated in new listening tests and the results of these and previous tests addressed by implementing respective *experts* in the Two!Ears model software environment. This first model is feature-based, reflecting the fact that *sound quality* of audio reproduction technology was found to result from features related to *coloration*, *spaciousness* and *artefacts* (Rumsey, 2002, Rumsey *et al.*, 2005).

In addition, tests on Quality of Experience evaluation and respective modelling work planned for year 3 are outlined in section 3.2 of Deliverable D 6.1.2. Where the most complex scenarios to be addressed by the final Two!Ears model are introduced, describing the functionality of the Two!Ears QoE model being aimed at.

This document is structured as follows: In the subsequent sections of this chapter, we outline the challenges encountered during the QoE model development in year 2 and the respective modelling decisions (Section 2.2). In Chapter 3, we present the results from the listening tests and measurements we carried out during the second year. Some of the data will then be modelled in Chapter 4.

2.2 Challenges and decisions on the way to model QoE

The goals of the Two!Ears QoE modelling activities and respective subjective test decisions were outlined in Deliverable D 6.2.1. The goal of the Two!Ears quality model is to tackle both sound quality related with a given spatial audio system and the *Quality of Experience* resulting for a listener. For sound quality evaluation[1], 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 situation is more complex, where the goal is to assess the overall listening experience (Raake and Blauert, 2013, Michael Schoeffler, 2013). In principle, this can be done by directly asking for the overall listening experience (Michael Schoeffler, 2013), where test settings such as the number of repetitions of a given audio content, the obvious variation of audio settings, or the parallel judgment of sound quality may direct the listeners' attention to the technical system. This consideration is related with the dual nature of multimedia perception, where humans can switch their attention between the content of the media in which she/he immerses, and the technical artefact that "transports" the respective information (see Mausfeld (2003)). More details on the assumed formation process of sound quality and QoE can be found in Raake and Egger (2014). Further direct assessment methods include asking about the liking of a given content and relating it to technical features, or to use content-related attributes to characterize the listening experience. Obviously, any such "guided" assessment (Jekosch, 2005) will have an impact on the result and not actually reveal the *Quality of Experience* of a person when listening to audio in an everyday listening situation ("Schrödinger's cat problem of QoE research", see Raake and Egger (2014)).

Because of these difficulties, it has been decided to focus the QoE assessment in Two!Ears on sound quality using separate coloration and localisation assessment (section 3.1 and 3.2) and *Quality of Experience* assessment using paired comparison on the one hand, and feature analysis with Multidimensional Scaling, see section 3.3.

The key challenges associated with the modelling plans in Two!Ears can be summarised as follows:

- The perceptual effects resulting from real-life spatial audio reproduction set-ups are rather small compared to degradations e.g. due to coding or low-cost electro-acoustic interfaces. As a consequence, test subjects tend to give rather high quality scores overall, or may not perceive large differences in the paired comparison tests.
- It is likely that there is no established reference in the minds of listeners when it comes

1 Sometimes referred to as "Basic Audio Quality" in the literature, in line with the terminology used in subjective quality test standards such as MUSHRA, BS.1534 (ITU) and respective models such as PEAQ (Thiede *et al.*, 2000).

to rather uncommon spatial audio reproduction systems such as the massive multi-channel Wave Field Synthesis (WFS). Instead, the best established reference most likely still is loudspeaker-based 2.0 or the less frequently used 5.1 stereophony. For these technologies, dedicated mixing paradigms and listening habits exist, which so far are not available for other spatial audio reproduction techniques. As a consequence, there is a strong impact of the source sequences used in the planned QoE tests in Two!Ears, requiring special consideration.

- For some of the planned *Quality of Experience* tests, it is unclear whether the assumed effects will actually be observed, for example in case of the impact of additional visual feedback. Hence, tests results will have to show whether the data will enable proper modelling.

- One of the most ambitious goals of Two!Ears is the linking of certain assessment results on *Quality of Experience* (not on sound quality) with features extracted using the Two!Ears model. Here, a great *audio experience* for listeners may not easily be explained based on the available bottom-up features or intermediate experts of the model.

Based on these challenges, a number of decisions have been taken, which are outlined in Deliverable D 6.2.1, namely:

- For sound quality, a feature-based model is being developed using coloration and localisation accuracy as the basis (Chapter 4). As ground-truth data, test results from MUSHRA-type coloration tests (for the original MUSHRA see ITU) and localisation tests are used. Here, too, considerations on the ability of identifying the number of sources have been made, linking the work in WP 6.2 with the planned modelling goals in WP 6.1 (Dynamic Auditory Scene Analysis).

- For QoE, the evaluation paradigm is two-fold: (1) Paired Comparison tests and multidimensional scaling (MDS) are employed to assess preferences between different audio reproduction set-ups. While this approach is still linked with an explicit evaluation of *sound quality*, the simple task enables to focus on what version of a presentation is perceived as better. The accompanying MDS addresses the underlying perceptual features that are related to certain preferences. (2) Indirect preference scaling using a method of construction is used in one case, where test subjects are to search for the preferred listening position in a given listening region, thus identifying the "sweet-spot" area, for different reproduction system configurations and contents.

- Modelling of sound quality is directly based on features extracted by the Two!Ears model.

- For QoE modelling, a mapping of differences in the auditory features for different stimuli to the corresponding listener ratings is planned, possibly assisted by an

intermediate mapping to the perceived features extracted via MDS. How exactly this part of the modelling will be addressed is still under investigation.

3 Contribution to Database in D 2.1

3.1 Coloration in Wave Field Synthesis

Wave Field Synthesis allows for the synthesis of a pre-defined sound field in an extended listening area, which is surrounded by loudspeakers. The limit in number of used loudspeakers leads to errors in the synthesized sound field. Those errors could have a negative effect on the ability to synthesize the desired spatial distribution of sound sources as well as on the sound color of the sound sources. As the errors occur only at higher frequencies – for most setups $> 1000\,\mathrm{Hz}$ – we could show that the perceptual influence is stronger on the perceived sound color than on the achievable localisation accuracy (Wierstorf *et al.*, 2014, Wierstorf, 2014).

Further investigation on the perceived coloration as presented in Wierstorf *et al.* (2014) showed that there were some numerical problems at very high frequencies in the used approach. Those problems most likely had influence on the perception of the listeners. We came up with a solution for the numerical problems by using a fractional delay (Laakso *et al.*, 1996) method in our simulations and rerun the listening test on coloration. The top row of Fig. 3.1 shows the results of the repeated listening test. The median together with the confidence interval is shown. Compared to the results of the first coloration experiment (Wierstorf *et al.*, 2014), a lower number of loudspeakers is now sufficient to avoid coloration in the synthesized sound field. But still, a loudspeaker spacing of 2 cm would be needed in a practical setup to achieve this.

The results from the top row of Fig. 3.1 were collected for a circular loudspeaker array with a diameter of 3 m. In addition to that loudspeaker array, we collected coloration ratings for a linear loudspeaker array with a length of 3 m, see the bottom row of Fig. 3.1.

The results are part of the Two!Ears database and D 2.2. They will be used in Chapter 4 to create and test a model for predicting the amount of perceived coloration.

The BRS (binaural room scanning files, which can directly be used with the Binaural Simulator of the Two!Ears model) files of this experiment are presented as database entry #36 in D 1.2, and the results as database entry #41 in D 1.2.

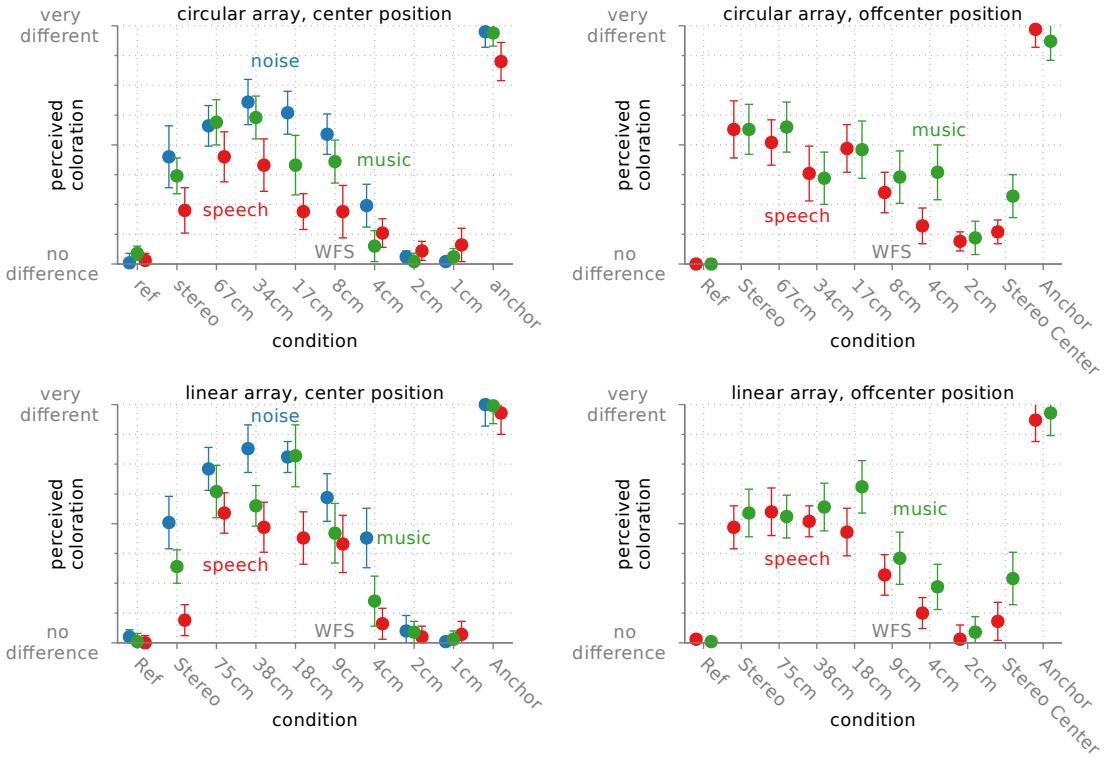

Figure 3.1: Coloration in WFS for a central and an off-center listening position. The median over 16 listeners together with the confidence interval is shown. For the WFS conditions different circular and linear loudspeaker arrays were applied, where the used loudspeaker distances are marked at the tics of the x-axes.

3.2 Coloration in Local Wave Field Synthesis

A second experiment was performed on the topic of coloration in WFS in close collaboration between TUB and URO. In this experiment we expanded the investigated sound field synthesis methods to include so called *local sound field synthesis* methods. The difference is that in this case the errors in the sound field are not distributed equally in the whole listening area as it was the case in the first experiment, but they can be avoided in one area and are more pronounced in other areas. The goal is then to create an area of the size of the human head inside the listening area where in the best case no perceptual coloration occurs. The experiment investigated for different local sound field synthesis methods if they are able to achieve this goal. One common local sound field synthesis method is band-limited Near-Field Compensated Higher Order Ambisonics (NFC-HOA), for which it is known that it creates a nearly artefact free region in the center of the listening area (Wierstorf, 2014). As Ambisonics is restricted to circular loudspeaker arrays,

the whole experiment was only conducted for a circular loudspeaker array. The other method investigated in this experiment, is so called Local Wave Field Synthesis (LWFS). It utilizes focused sources as virtual loudspeakers around the head of the listener, which are then individually driven by WFS to create the desired sound field (Winter and Spors, 2015). This shrinking of the listening area is similar to the spatial band-limitation exploited in band-limited Near-Field Compensated Higher Order Ambisonics. In both cases, the shrinking will reduce the perceptual errors in the synthesized sound field in the given small area.

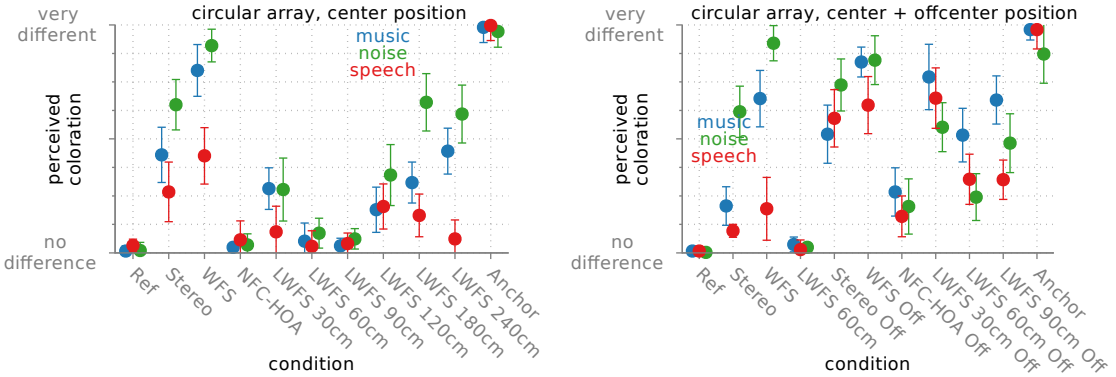

Figure 3.2: Coloration in local sound field synthesis for a central and an off-center listening position. The median over 17 listeners together with the confidence interval is shown.

Figure 3.2 summarizes the results for two different runs of the experiment. In one, only a central listening condition was considered (left graph). The other run of the experiment compared a central listening position for some reproduction systems with an off-center listening position for other or the same reproduction systems. First, we will discuss the results for the central listening position only. The main result is that NFC-HOA and LWFS with a diameter of 60 cm and 90 cm of the local area is not significant different from the reference condition, which was a single loudspeaker as in the tests before. This shows that both techniques, NFC-HOA and LWFS are able to generate a small area that has no perceptual change in timbre compared to a given reference. If the local area is enlarged for LWFS a clear change in timbre is observed. This change is stronger for music and noise compared to speech as source material.

In the right graph the notation *Off* denotes listening conditions, where the listener was positioned 1 metre left from the center of the loudspeaker array. All conditions without *Off* correspond to the central listening position are equivalent to the respective pendant in the left graph. For the major number of reproduction methods, namely Stereo, WFS and NFC-HOA no adjustments depending on the listener position have been applied. For the LWFS the driving functions were modified such that the local listening area is always centered at the listener position.

Interestingly, all of the LWFS conditions are now perceived to be more colored compared to the reference as it is the case for the NFC-HOA condition. In future experiments we will investigate what might be the main reason behind this, as there are several differences between LWFS and band-limited NFC-HOA.

The BRS files of this experiment are presented as database entry #35 in D 1.2, and the results as database entry #40 in D 1.2.

3.3 Sound quality in Wave Field Synthesis

The coloration ratings presented in the previous sections only provide a distance metric from the given reference. As the timbral space is multi-dimensional it cannot be stated if two stimuli rated to have the same coloration regarding the reference sound similar or not. This implies that we can also not conclude directly from the coloration rating to the perceived sound quality of the presented stimuli. Let us assume that the only difference in the perception of the stimuli is indeed the coloration, even then we cannot conclude that two stimuli rated to have the same coloration would have also the same sound quality rating.

To investigate this further we used the same stimuli as in the coloration experiment presented in Section 3.1. We conducted two experiments with it, in the first one listeners were asked to judge the preferred sound quality in a paired comparison paradigm. In the second experiment they were asked to judge the perceptual difference between presented pairs. From the first experiment we can create an ordering of the stimuli regarding their perceived sound quality. From the second experiment we have distance ratings between the different stimuli which can be used in a multi-dimensional scaling analysis to create a perceptual space and relate the coloration and sound quality ratings to this space.

For both experiments, only the results of the single listeners are available in the database as the analysis of the data will be performed after this Deliverable.

The BRS files of this experiment are presented as database entry #36 in D 1.2 as they are the same as in the coloration experiment, and the results as database entry #42 and #43 in D 1.2.

3.4 Binaural room impulse responses for a 5.0 surround setup

We are preparing an experiment, in which listeners should find the sweet-spot position for listening to different 5.0 surround recordings under different amount of visual information on the presented scene. For the purpose of modeling the results of this experiment we need the ear signal of the listener at the different listening positions in order to decide which position is the best one. To allow for this we decided to do the experiment with the help of dynamic binaural synthesis, which has the advantage that the listener and the model will listen to exactly the same physical stimuli. For the dynamic binaural synthesis we need binaural room impulse recordings at different listening positions. We decided to record those at nine different positions in a 5.0 loudspeaker setup in a studio room. The data is provided as database entry #39, see D 1.2.

4 Model implementation

In this section, we describe the actual work on the implementation of the QoE relevant parts of the Two!Ears model. As the results from the listening tests on different spatial audio systems are mainly available for the two attributes direction and coloration of a sound source, the model implementation was focused on those attributes as well. The prediction of sound quality ratings, incorporation of pre-knowledge, and more advanced adjustment of the inner reference will follow in the third year.

4.1 Predicting the direction of an auditory event

For different spatial audio systems the ability to synthesize a point source placed at a particular position depends on the amount of applied loudspeaker and the position of the listener within that system. For example, for stereophonic systems there exists only a small area in which the spatial impression is correct, the so called sweet-spot. For sound field synthesis methods this area becomes larger and the localisation of a synthesized point source can be undistinguishable from a real one, see the results in Wierstorf (2014) and database entry #26 to #31 in the Two!Ears database (D 1.1).

The goal is to model the localisation abilities in different sound field synthesis systems to be able to include them in the final quality model, as one spatial attribute that could have an influence on the perceived QoE.

4.1.1 Predicting the direction

From a modeling perspective the task is challenging, as the physical signals contain lots of artefacts above a given frequency (which could range from 100 Hz up to 1300 Hz for typical setups). Those artefacts could lead to contradicting binaural features that are normally used for localisation such as interaural time (ITD) and interaural level differences (ILD). The longterm goal is to use a common localisation stage in the Two!Ears model that can cope with the sound field synthesis stimuli as well as with the localisation tasks in complex environments, like a room with lots of reverberation and competing sources. We showed already that multi-conditional training is a possible way to achieve this (May *et al.*, 2015).

As this is currently not available in the TWO!EARS model we restricted us in the first version to a simple ITD-azimuth lookup table. This has been shown to provide reasonable good predictions (Wierstorf, 2014).

The implementation is done as the *ItdLocationKS* in the blackboard. All the results from the listening tests were modelled and summarised in Fig. 4.1 to Fig. 4.6. The modeling is also explained in the official TWO!EARS documentation as an example[1].

4.1.2 Prediction results

The model performance is compared to the results from the implementation presented in Wierstorf (2014). There, the binaural model after Dietz *et al.* (2011) was used in combination with a lookup table and some outlier detection to predict the perceived direction. The current TWO!EARS implementation uses the same lookup table and outlier detection mechanism, but the ITD cues are provided by the Auditory Front-End of the TWO!EARS model which extracts them in a different way than the model after Dietz *et al.* (2011) does it. Nonetheless the results of both modeling approaches are very similar. For most of the conditions, the TWO!EARS model predicts the perceived directions slightly better. Only for the case of the synthesized focused source, the TWO!EARS model localises the synthesized focused source better than the human listeners which leads to larger prediction errors.

1 `http://twoears.aipa.tu-berlin.de/doc/latest/examples/qoe-localisation`

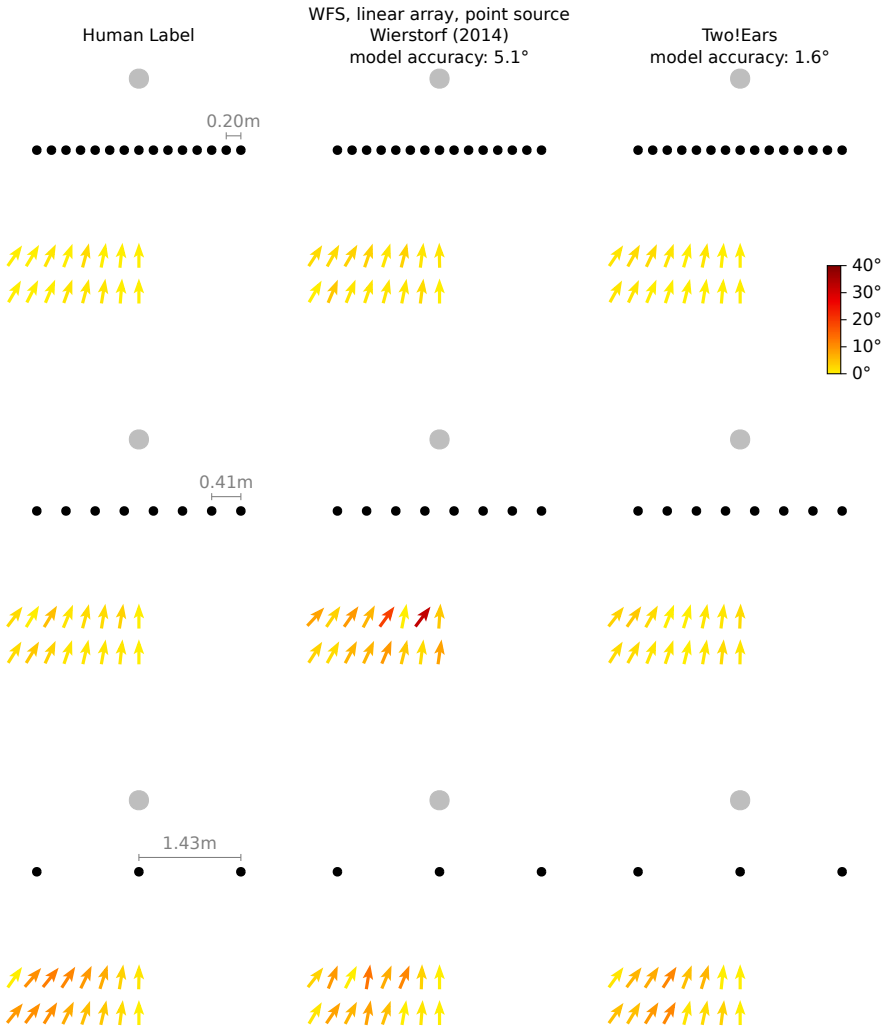

Figure 4.1: Average localization results and predictions. The black symbols indicate loudspeakers, the grey ones the synthesized source. On every listening position an arrow is pointing into the direction the listener perceived the corresponding auditory event from. The color of the arrow displays the absolute localization error. The model predictions are shown in the center column for the modeling approach after Wierstorf (2014) and for the Two!Ears model in the right column. The model accuracy is given as an average over all listener positions and loudspeaker setups.

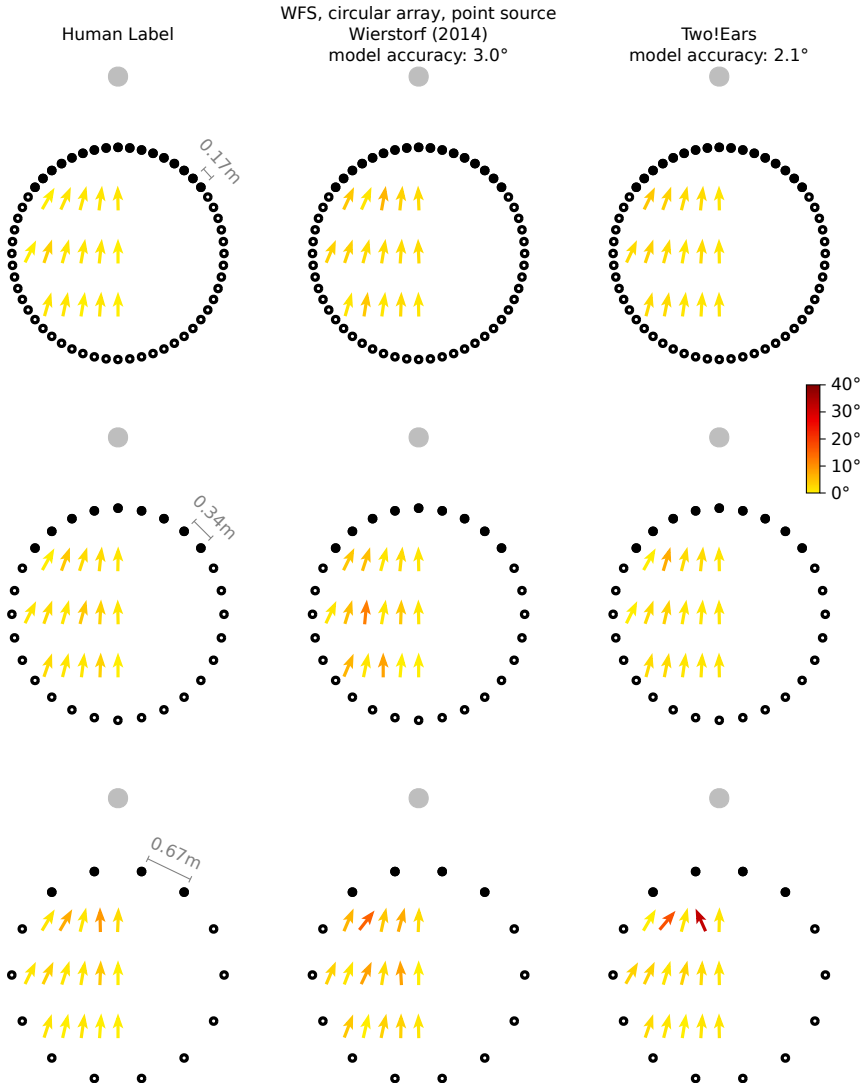

Figure 4.2: Average localization results and predictions. The black symbols indicate loudspeakers, the grey ones the synthesized source. On every listening position an arrow is pointing into the direction the listener perceived the corresponding auditory event from. The color of the arrow displays the absolute localization error. The model predictions are shown in the center column for the modeling approach after Wierstorf (2014) and for the Two!Ears model in the right column. The model accuracy is given as an average over all listener positions and loudspeaker setups.

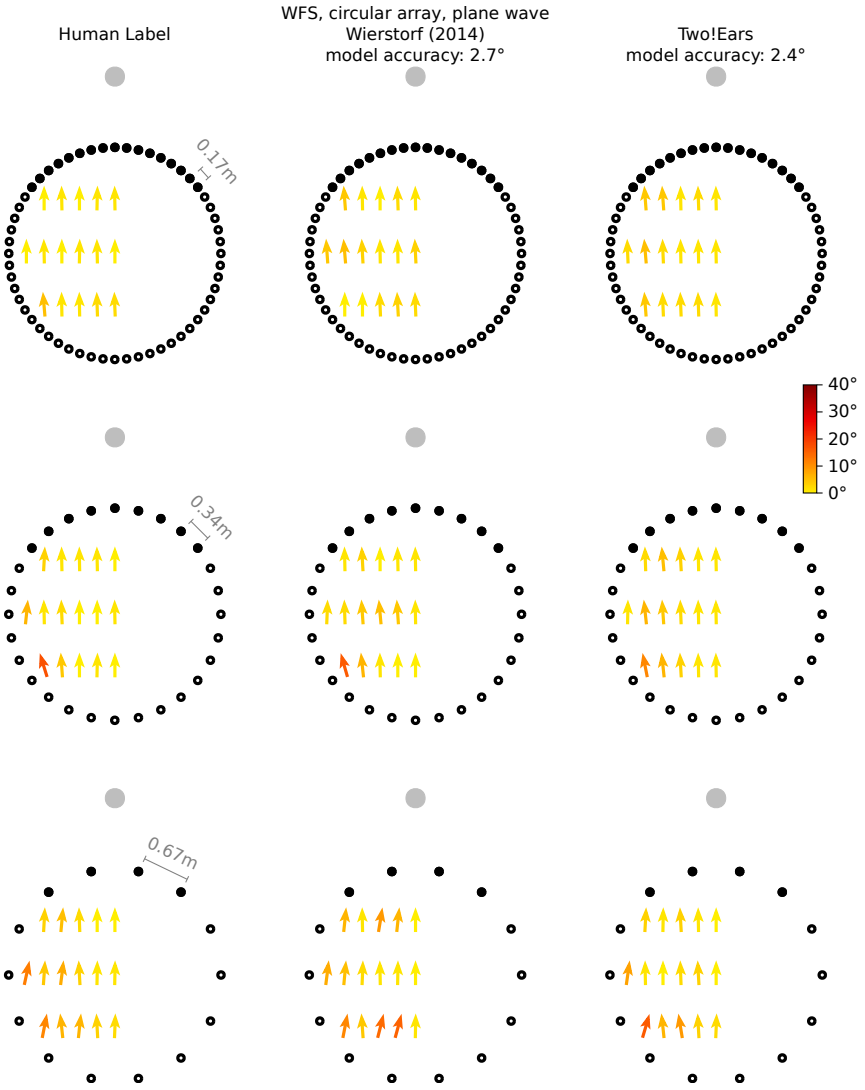

Figure 4.3: Average localization results and predictions. The black symbols indicate loudspeakers, the grey ones the synthesized source. On every listening position an arrow is pointing into the direction the listener perceived the corresponding auditory event from. The color of the arrow displays the absolute localization error. The model predictions are shown in the center column for the modeling approach after Wierstorf (2014) and for the Two!Ears model in the right column. The model accuracy is given as an average over all listener positions and loudspeaker setups.

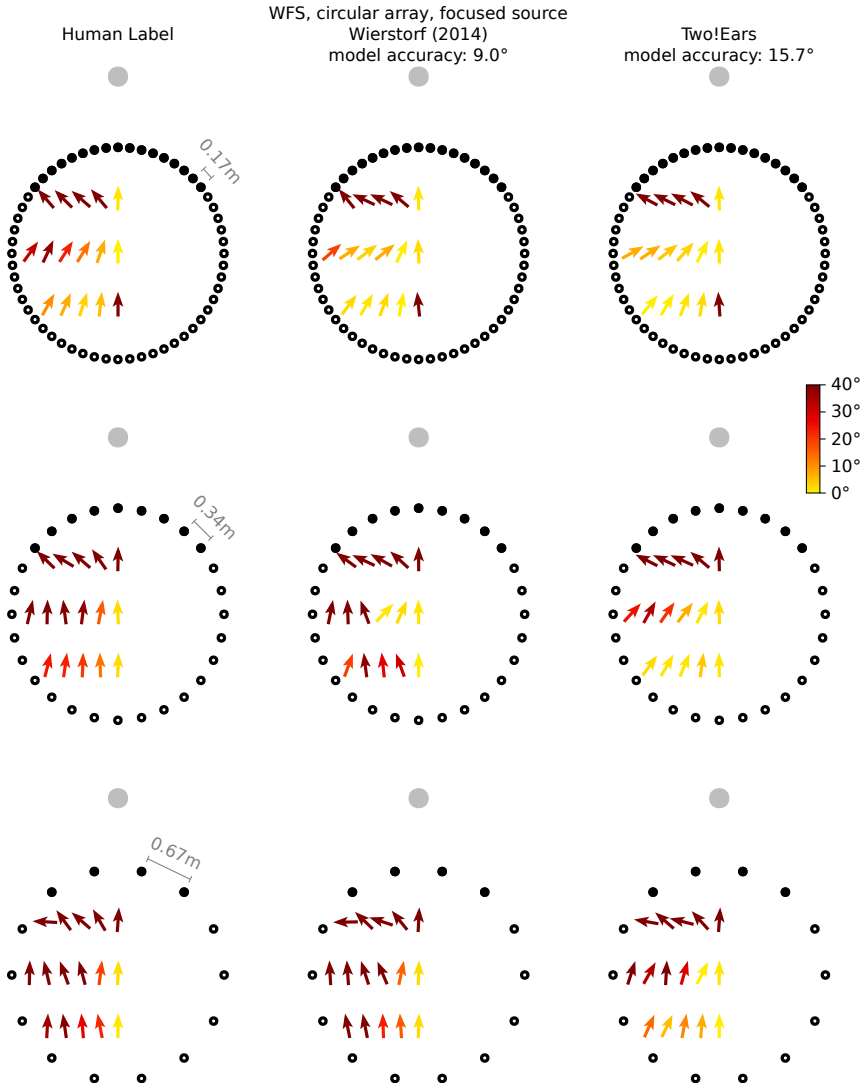

Figure 4.4: Average localization results and predictions. The black symbols indicate loudspeakers, the grey ones the synthesized source. On every listening position an arrow is pointing into the direction the listener perceived the corresponding auditory event from. The color of the arrow displays the absolute localization error. The model predictions are shown in the center column for the modeling approach after Wierstorf (2014) and for the Two!Ears model in the right column. The model accuracy is given as an average over all listener positions and loudspeaker setups.

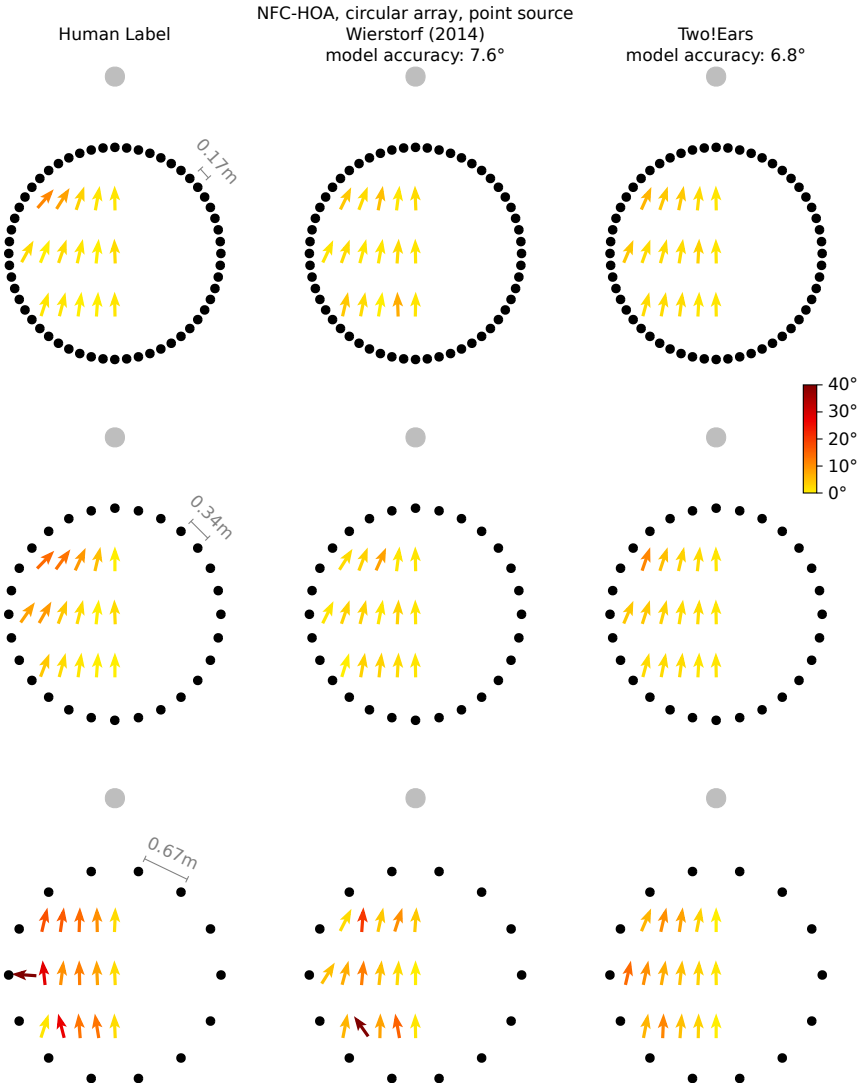

Figure 4.5: Average localization results and predictions. The black symbols indicate loudspeakers, the grey ones the synthesized source. On every listening position an arrow is pointing into the direction the listener perceived the corresponding auditory event from. The color of the arrow displays the absolute localization error. The model predictions are shown in the center column for the modeling approach after Wierstorf (2014) and for the Two!Ears model in the right column. The model accuracy is given as an average over all listener positions and loudspeaker setups.

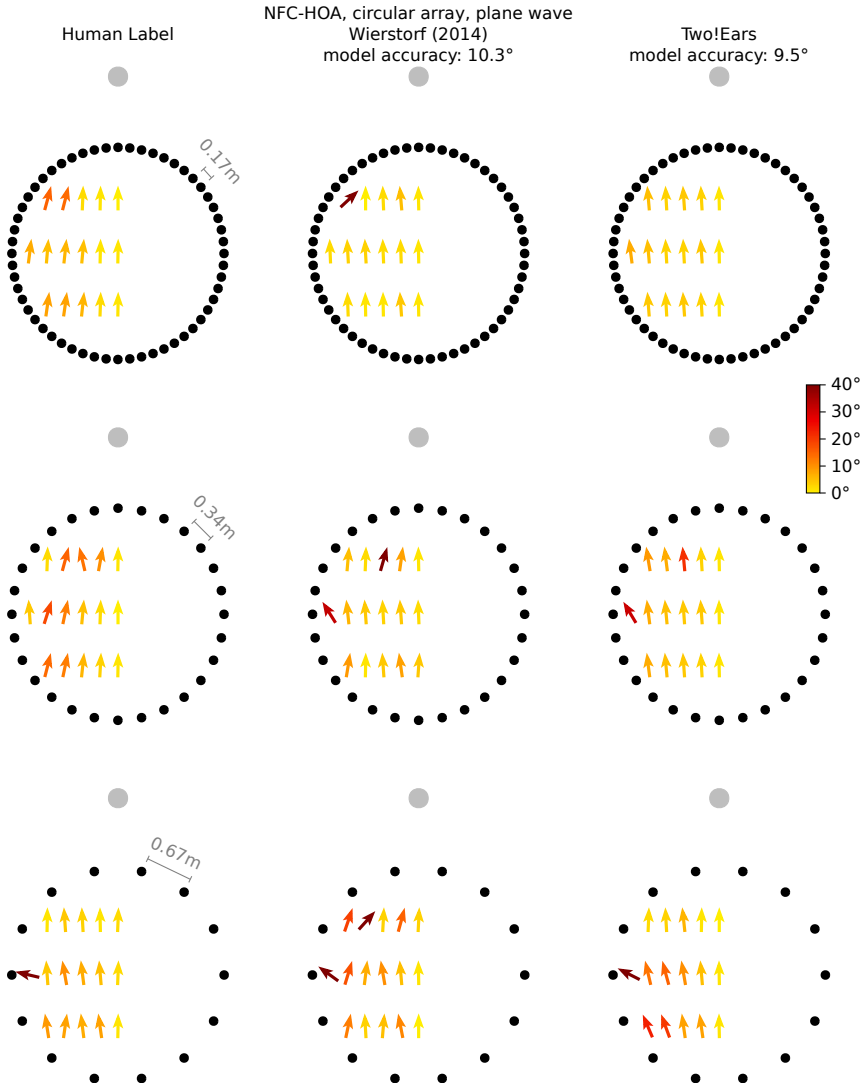

Figure 4.6: Average localization results and predictions. The black symbols indicate loudspeakers, the grey ones the synthesized source. On every listening position an arrow is pointing into the direction the listener perceived the corresponding auditory event from. The color of the arrow displays the absolute localization error. The model predictions are shown in the center column for the modeling approach after Wierstorf (2014) and for the Two!Ears model in the right column. The model accuracy is given as an average over all listener positions and loudspeaker setups.

4.2 Predicting the coloration of an auditory event

The model prediction for coloration of a sound source is more difficult as the prediction of its perceived direction. There are several factors adding up to this difficulty. First, coloration describes a change in timbre from one point in the timbral space to another one. This means we start always from a so called reference point (in the experiment labeled as the reference stimulus) to which the listeners should compare the timbral perception of another test stimulus. We could also directly ask for coloration of a presented stimulus without presenting the reference, but this does not mean the listeners are using no reference, but that they use a learned reference for this particular situation. Another problem comes with the fact that the timbral space is multi-dimensional and the position in its space depend on several signal features, which could be most noticeable in the frequency-spectrum of the stimulus or in its time-domain.

In order to create a coloration knowledge source in the Two!Ears model we will narrow the problem. As the sound quality of spatial sound systems are the main application of the model in D 6.2 we focus the coloration modeling on those stimuli. In spatial audio systems the most pronounced signal features that correlate with a change in timbre are comb-filter like artefacts in the frequency spectrum of the signals, as different loudspeaker signals sum up at the listener position, compare Fig. 5.8 in Wierstorf (2014). This simplifies the prediction of coloration by the fact that we can focus on spectral auditory features only, namely the output of the gammatone filterbank of the Two!Ears Auditory Front-End.

4.2.1 Predicting coloration

The implementation in the Two!Ears model is done in the form of a *ColorationKS* (knowledge source) in its blackboard system. For the prediction of the coloration of a synthesized sound source we used the model proposed by Moore and Tan (2004). In the original paper the authors used it to predict the naturalness of different comb-filtered stimuli. As this was the only factor they changed in their stimuli it is very likely that their listener rated naturalness in the same way they would have rated coloration for those stimuli.

The basic idea of their model is to compare the two weighted excitation patterns of a test stimulus and a reference stimulus. The excitation patterns are calculated by a gammatone filterbank and after that the standard deviations across the frequency channels of the differences between the two excitation patterns are calculated. The standard deviation is calculated for the direct differences between both excitation pattern and for the differences between their slopes. The final difference value is then a weighted sum of both standard deviations.

The model has two different sets of parameters for speech and noise/music as stimuli. We use those settings as well, by informing the ColorationKS which type of stimuli it will listen to. Note, that this could be done in a later state also by a classification knowledge source. In the original model only pink noise was used for the prediction, even in the case of speech stimuli in the experiment. As we would prefer to have the prediction for all type of stimuli we will not use this restriction. As this will introduce the possibility of non-stationary stimuli, we will reexamine during the next steps if we will get better predictions if we use time-varying excitation patterns and not time averaged ones as we do at the moment.

As the model requires the excitation pattern of the reference, this has to be known by the ColorationKS as well. We decided to implement it in the storage of the blackboard system. This implies that it could be easily learned and also changed and adjusted by other knowledge sources. For example, if we would like to add the ability to change the internal reference in a context dependent manner.

4.2.2 Learning the reference

The learning of the reference is implemented in an automatically way at the moment. The memory of the blackboard system is unique to one instance of it. If we initialize a new blackboard the learned reference will be empty. In this case, the blackboard will calculate the auditory features from the first signal it is presented with and stores the result as the new reference. All other incoming signals will then compared to this reference.

For an practical example, imagine a MUSHRA listening experiment. For every run of the experiment we would initialise a new blackboard, present first the reference signal to the model, and after that all the test stimuli.

4.2.3 Prediction results

We applied the coloration part of the Two!Ears model to the listening test results we obtained for different sound field synthesis methods (see section 3.1). As the Binaural Simulator of the Two!Ears model is able to directly handle binaural room scanning files that are normally used for the binaural simulations of spatial audio systems, we can fed the exact same stimuli into the model as we used during the listening test (see database entry #36 in D 1.2). The audio material of the listening test consisted of speech, pink noise and music, with a length of around 9 s. For the modeling we limited the length of all stimuli to the first 5 s.

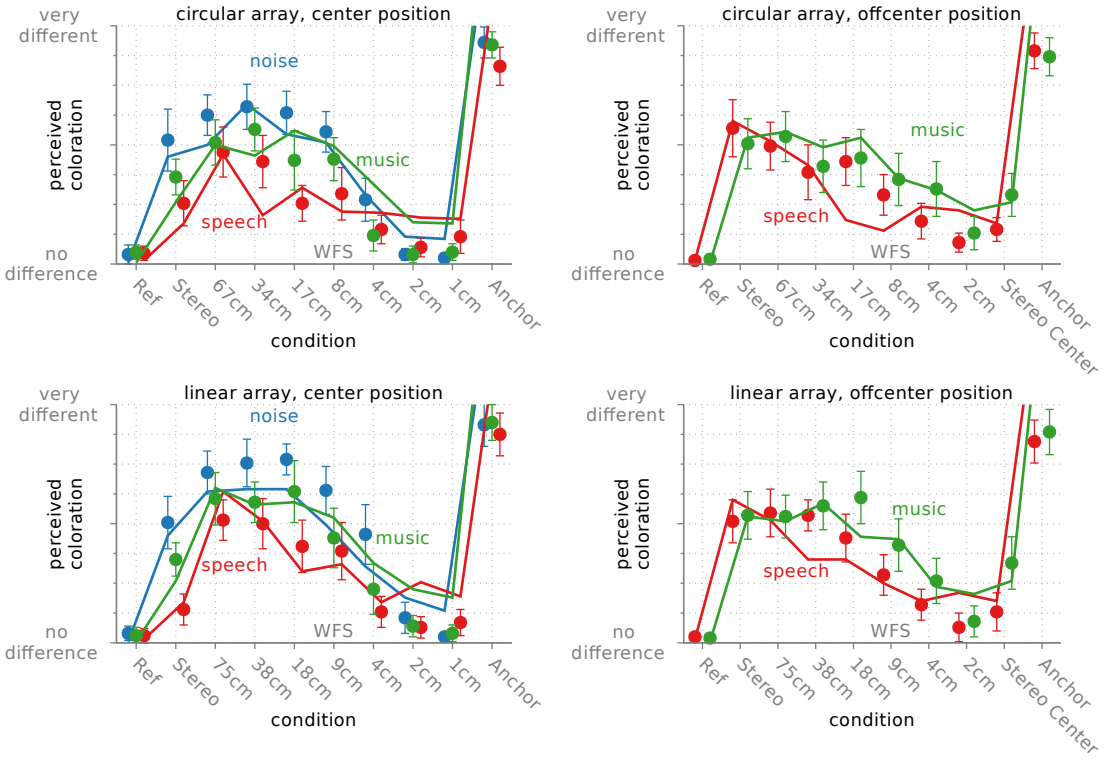

Figure 4.7: Coloration in WFS for a central and an off-center listening position. The median over 16 listeners together with the confidence interval is shown (points) together with the model predictions (lines). For the WFS conditions different circular and linear loudspeaker arrays were applied, where the used loudspeaker distances are marked at the tics of the x-axes.

Figure 4.7 presents the results of the model. The standard parameter as explained on page 906 in Moore and Tan (2004) were used. The only adjustment was a scaling of the resulting difference values to fit in the same range as the listening results. This was done by the same value for all conditions and audio source materials. There are a few points where the model prediction is significantly different from the listening test results, but overall it is in good agreement with the results. The model is able to predict the difference in coloration depending on the input signal, which is a desirable output as the original model from Moore and Tan (2004) was only designed to do all its predictions by the usage of pink noise.

The coloration model with exactly the same parameters was further applied to the listening test results for Local Wave Field Synthesis, see section 3.2. Figure 4.8 summarizes the results. The model has some problems to predict larger coloration values with high accuracy, but it is able to identify which techniques provide more or less no coloration and which techniques suffer from larger changes in timbre.

In a next stage the model will be analysed in a detailed fashion for those conditions where it fails, to see how it can be improved. In addition, it will be tested if a time varying coloration prediction will enhance the results regarding different audio source materials.

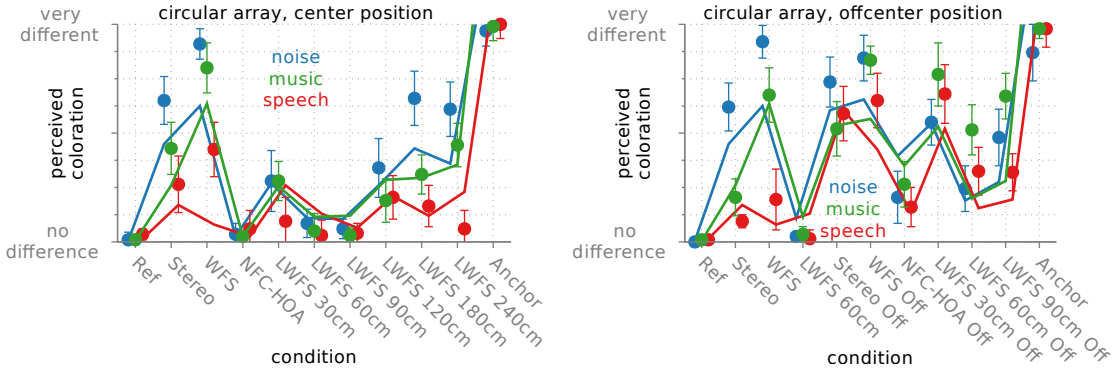

Figure 4.8: Coloration in LWFS for a central and an off-center listening position. The median over 16 listeners together with the confidence interval is shown (points) together with the model predictions (lines). For the WFS conditions different circular and linear loudspeaker arrays were applied, where the used loudspeaker distances are marked at the tics of the x-axes.

5 Conclusions

We discussed the current status of the part of the TWO!EARS model that will be applied to the prediction of aspects of *Quality of Experience* in spatial audio systems. The listening test results started with assessment of basic attributes of sound quality like coloration and localisation. The first version of the *Quality of Experience* model also focussed on those parts.

We presented results from further listening tests that where needed for modeling coloration and binaural measurements as a preparation for further tests on sound quality.

The current state of the TWO!EARS model is able to predict coloration and localisation for multiple spatial audio systems and different listener positions. The localisation will be further improved with the upcoming version of the model, using more robust localisation stages and the ability to detect the number of sound sources.

The most challenging part in the third year will be to identify features beside localisation and coloration in the ear signals of the listeners that will allow the prediction of the rated *Quality of Experience*.

Bibliography

(), *ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems*, International Telecommunications Union. (Cited on pages 4 and 5)

Dietz, M., Ewert, S. D., and Hohmann, V. (**2011**), "Auditory model based direction estimation of concurrent speakers from binaural signals," *SpeechCom* **53**(5), pp. 592–605. (Cited on page 14)

Jekosch, U. (**2005**), *Voice and Speech Quality Perception — Assessment and Evaluation*, Springer, D–Berlin. (Cited on page 4)

Laakso, Valimaki, Karjalainen, and Laine (**1996**), "Spliting The Unit Delay," *IEEE Signal Processing Magazine* **1**(January), pp. 30–60. (Cited on page 7)

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)

May, T., Ma, N., and Brown, G. J. (**2015**), "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *ICASSP*. (Cited on page 13)

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 page 4)

Moore, B. C. J. and Tan, C.-t. (**2004**), "Development and Validation of a Method for Predicting the Perceived Naturalness of Sounds Subjected to Spectral Distortion," *Journal of the Audio Engineering Society* **52**(9), pp. 900–14. (Cited on pages 21 and 23)

Raake, A. and Blauert, J. (**2013**), "Comprehensive modeling of the formation process of sound-quality," in *Proc. IEEE QoMEX*, Klagenfurt, Austria. (Cited on page 4)

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. (Cited on pages 3 and 4)

Rumsey, F. (**2002**), "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 page 3)

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 3)

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 page 4)

Wierstorf, H. (**2014**), "Perceptual Assessment of Sound Field Synthesis," Ph.D. thesis, TU Berlin. (Cited on pages 7, 8, 13, 14, 15, 16, 17, 18, 19, 20, and 21)

Wierstorf, H., Hohnerlein, C., Spors, S., and Raake, A. (**2014**), "Coloration in Wave Field Synthesis," in *AESC55*, pp. Paper 5–3. (Cited on page 7)

Winter, F. and Spors, S. (**2015**), "Physical Properties of Local Wave Field Synthesis using Circular Loudspeaker Arrays," in *EuroNoise*, Maastricht, The Netherlands. (Cited on page 9)