A case for TWO!EARS in audio quality assessment

Alexander Raake, Hagen Wierstorf
Assessment of IP-based Applications, Technical University Berlin, Germany.

Jens Blauert
Institute of Communication Acoustics, Ruhr-University Bochum, Germany

Summary

The paper presents first results on audio-quality evaluation following the auditory-perception modeling paradigm of TWO!EARS, a FET-Open project funded under FP7 ICT by the European Commission (`www.twoears.eu`). The project targets an interactive system for binaural auditory perception, including input from the visual modality. In its ultimate form, it will be based on an interactive robot platform combining bottom-up cues from monaural and binaural signal processing with hypothesis-driven top-down cognition, organized around an expert-system with black-board architecture. One of the two proof-of-concept applications is that of audio-quality evaluation for loudspeaker-based reproduction. Here, targeted key innovations include a dedicated scene-based evaluation paradigm, active exploration of sound fields using head-movements and displacement, the combination of bottom-up information with top-down feedback and adaptation, and the dedicated use of internal rather than external references within the expert system. The paper presents research conducted in the course of model development. Two first findings along these lines will be presented: (1) A listening test was conducted to assess the localization performance of human listeners for several near-field compensated higher-order Ambisonics setups. A binaural model was able to predict the perceived direction with good accuracy. However, for some positions in the listening area, listeners reported to perceive more than one auditory event. This result was not predicted by the binaural model, which instead returned a single source direction. By endowing the binaural model with the ability of head rotations, it could be expanded so as to more accurately predict whether the listener perceived one or more sources, and from which direction. (2) A first pilot test has been conducted to address the scene-based test-paradigm targeted by TWO!EARS. Here, a scene with two guitars and one singer playing a musical piece was reproduced via different Wave Field Synthesis systems so as to selectively degrade one or multiple of the three sources. The different conditions were assessed in a paired comparison preference test, indicating, among other findings, that the reference scene was clearly not the preferred one. The paper summarizes the results and provides an outlook on future developments on sound quality modeling in the TWO!EARS project.

PACS no. 43.66.Ba, 43.60.Sx

1. Introduction

In [1] we have introduced a novel sound quality model framework. It represents the sound quality assessment application of the multi-modal perception model currently developed in the project TWO!EARS (`www.twoears.eu`). The TWO!EARS model targets the integrative modelling of the bottom-up auditory signal processing, human cognition and top-down processing, including aspects of audiovisual interaction. The system is centered around a blackboard-structure that is built on a graphical model architecture [2, 3].

When audio signals are played back via loudspeakers or headphones, different elements of the end-to-end chain from creation to presentation may impact the sound quality perceived by listeners. Innovation of the past years has addressed all of these parts, for example in terms of new multi-microphone recording and processing approaches, new multichannel representation and coding techniques [4, 5, 6], and especially sound field reproduction systems that enable the presentation of spatial audio [7]. The ultimate users of such audio systems are listeners. Hence, for all elements of the audio production, delivery and presentation chain, the primary performance-criterion for the technical system design is the quality the listener perceives.

Obviously, like all perceptual events, quality 'happens' in the brain of the listener [8]. 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 of quality has been coined as *quality (based on experiencing)* [9]. 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/he considers to purchase, or when she/he is a test participant in an audio quality listening test.

When the auditory listening experience at large is evaluated, this can be referred to as *quality of experience*, which has been defined as (see [9], extending [10]):

> *Quality of Experience (QoE)* 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.

It is noted that in this case the listener (person) does not necessarily need to be aware of how the underlying technology influences the listening experience. It is obvious that real QoE according to this definition is hard to assess, and in practice, most research labeled as QoE-research is rather concerned with quality assessment based on experiencing. The same holds for the present paper, where the participants of the underlying tests are aware that they are assessing a technical system. In the following, we will refer to this type of quality assessment in terms of *sound quality.* For an overview of sound quality evaluation see, for example, Bech and Zacharov [11].

In this paper, we present the results of two pilot experiments to underline the usefulness of the proposed model paradigm:

1. Based on data from a localisation test described in [12], it is shown that head-movements implemented in a binaural model (modified from [13]) may improve the localisation predictions and bring it closer to the localisation data obtained from human listeners.
2. In a paired-comparison preference test, a musical piece with three sound sources (two guitars and a singer) reproduced over different audio reproduction systems was assessed. Here, it was shown that the source file, as it is normally used for reference in sound quality tests, was not the preferred choice. Instead, certain degradations of specific sources were found to be preferred by the test listeners. This study highlights the usefulness of the scene-specific paradigm. Moreover, the paper informs about the usefulness of alternative test methods that enable a holistic sound quality evaluation.

The paper is structured as follows: Sec. 2 provides an overview of sound quality research, Sec. 3 summarizes considerations on sound quality assessment relevant for the two experiments presented in In Sec. 4. Sec. 6 provides conclusions and an outlook to next steps in TWO!EARS research.

2. Sound quality assessment

The key aspect in sound quality evaluation of spatial audio is the fact that humans perceive acoustic scenes based on an auditory scene analysis. Based on the 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 [8]. Now, when the sound quality of a reproduction system is assessed in terms of a comparison of perceived features with regard to desired features, an unanswered research-question is what kind of internal reference is being used by the listener [8, 14]?

2.1. Sound quality dimensions & assessment

For the systems addressed in this paper, namely multi-channel loudspeaker sound field synthesis, headphone-based binaural synthesis or stereophonic systems, quality is determined primarily by timbral and spatial features, and by spectro-temporal artifacts [15, 16, 17, 18, 19, 7]. For stereophonic sound reproduction evaluation targeting 5.1 systems and using a scene-based paradigm [15], Rumsey et al. found quality to be determined by *timbral* and *spatial fidelity*, which explained 70% vs. 30% of the quality variance, respectively [17]. Extensive spatial fidelity assessment has been conducted, for example in [16, 17, 18, 19, 7]. The timbre related with different WFS systems has been studied in Wierstorf et al. [20]. Artifacts may, for example, be introduced by spatial aliasing as it occurs in practically realized multichannel-audio presentation [21]. It should be noted, that added artifacts may lead to additional auditory streams processed separately from the underlying scene [22].

Often times, explicit reference stimuli are used in quality tests, for example in the tests typically conducted for audio coding quality evaluation. Known methods of this type are MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor, [23]) and BS.1116 [24]. 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 [25, 26]. Here, specific stimuli are often presented as hidden references that are not identified as such to the test participants.

2.2. Instrumental methods for sound quality assessment

In many cases, auditory quality tests with subjects are too time-consuming to be carried out, or the system to be tested still is in the design phase and hence not fully implemented. Here, instrumental methods are often used instead. Two fundamental types of methods can be distinguished:

1. Algorithms or metrics that are based on physical properties of the signal or sound field, which may be put into relation with perceptual attributes or ratings.
2. Algorithms that implement specific parts of the auditory signal processing, possibly including cognition-type mapping to quality dimensions or overall quality.

An examples of a type-(1) measure for the case of sound field synthesis addresses the deviation of the reproduced sound field from the desired one. An example for room acoustics evaluation metrics are reverberation decay times.

Although such measures enable a direct relation with the physical properties of the system, they do not capture the actual perception by listeners. Here, the *explicit modeling* of human signal processing in terms of type (2) and respective mapping to quality ratings resulting from listening tests provide a more valid quality evaluation. Different elaborate approaches of this type have been developed in the past years, and have been standardized in bodies such as the International Telecommunication Union. Examples include PESQ [27] and POLQA [28, 29] for speech transmission systems, and PEAQ [22] for audio coding evaluation. These so-called signal-based, full-reference models estimate quality comparing the processed audio signal with an unprocessed reference.

For instrumental sound reproduction assessment, first models have mainly been based on the framework of spatial and timbral fidelity (e.g., [17]). To this aim, underlying technical or physical characteristics of the acoustic scene are mapped to low-level attributes or perceptive constructs. For example in QESTRAL [30], spatial fidelity is predicted from perceptually relevant cues such as inter-aural time and level differences. Some approaches have been proposed for timbral fidelity prediction: [31] describe a model for coloration prediction of bandpass-filtered speech and audio. Another coloration-prediction model for room acoustics is presented in [32].

Full-reference models for audio reproduction are currently developed by ITU-R SG6 [33], and different algorithms have recently been described in the literature [34, 35]. As for coding-related FR-models, to derive a quality estimation, the processed and reference signals are first analysed in terms of model output variables (MOVs), for example by models of the auditory periphery, reflecting certain aspects of human auditory bottom-up processing. In a subsequent step, aspects of human cognition are applied, for example targeting a relevance-weighting of different MOVs (see [22, 34, 35] for examples).

The current quality models have different limitations (see [1] for more details): Explicit references are typically used, and internal references are considered only partially (e.g. [29, 35, 34]), the scene-based approach taken by some few models is rather rudimentary (foreground/background considerations in [30], some implicit aspects in [22, 35]), active exploration is currently unavailable in models, and the peripheral models currently used may be complemented by further features, and specifically be enhanced by top-down feedback [2].

We attempt to address these aspects with the TWO!EARS sound quality model. The general modeling paradigm has been outlined in [1]. It is based on a comprehensive model of interactive listening and auditory-scene analysis [36, 2]. The modular architecture of the model targets to functionally replicate human hearing and relevant aspects of cognition. The targeted sound-quality model has the following particular properties (adopted from [1]; the aspects partially addressed in this paper are marked with arrows):

⇛ 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 (for example room-acoustics- vs. sound-source-quality, different usage scenarios for the evaluated room, etc.).

- 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 experts 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 [37].

- 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 as a function of different media representation approaches (e.g. [38, 39]), attempts will be made to extend the paradigm from sound quality to actual QoE modelling.

3. Methodological aspects addressed in this paper

In this section, different considerations regarding sound quality assessment relevant to the design of the two pilot studies are summarized.

3.1. Interactive localisation prediction

The majority of tests conducted for sound quality evaluation of loudspeaker-based sound field synthesis follows the "spatial and timbral fidelity" paradigm outlined above [30]. In cases of real-life usage, there mostly is no explicit *"original"*, and judgments of quality result from a comparison of the perceived character of the scene with internal instead of explicit external references. Real-life audio stimuli listened to by users of audio systems have typically been generated in the course of a creative process involving musicians or speakers, or other types of audio generation, subsequent recording and signal processing, as well as audio reproduction. Here, listening habits determine the world-knowledge of the listener and the respective internal references applied during quality evaluation [9, 39]. These considerations limit the usefulness of the "fidelity" paradigms previously adopted in sound quality research for stereophonic audio reproduction.

Our work in TWO!EARS will ultimately target evaluations in relation to internal rather than external explicit references. It is undisputed, however, that sound quality of spatial audio reproduction is of multidimensional nature, and that two key compounds of quality dimensions are that of timbre or coloration, and that of the spatial character of the delivered scene. Besides sound quality evaluation, another application scenario addressed by TWO!EARS is that of interactive auditory scene analysis, for example in a search-and-rescue context. Here, localisation of sound sources is a key-functionality. A first implementation of the interactive model based on a blackboard system built using a Bayesian graphical model architecture has been applied to interactive localisation in [3], showing that head-movements can significantly improve localisation performance.

From our previous work on sound field synthesis evaluation, there is a large body of localisation test data from human listeners available [40, 12]. In that work, the human localisation data has also been compared to predictions from binaural models such as the Dietz model [13]. Now, it is a straight-forward approach to extend the modeling paradigm by considering active exploration by the binaural model, and investigate whether the predictions get closer to the ones obtained from listeners in the localisation tests. This approach represents the first pilot test on the TWO!EARS sound quality evaluation paradigm.

3.2. Scene-related sound quality test

The specific question addressed with this second experiment is that of a scene-specific evaluation: Does sound quality depend on which of three different sources in a given scene are degraded by specific processing? Details about the test set-up are outlined in Sec. 4.

Instead of tests with an explicit reference, single-stimulus tests such as Absolute Category Rating with, for example, a 5-point scale are often used [25, 26]. Such tests are assumed to better focus the participants' evaluation criteria on a comparison with internal rather than external, explicit references. In practice, however, this test paradigm has other limitations. The biggest limitation is that smaller quality-differences cannot easily be assessed with this method, without the requirement for a large number of test participants. Obviously, the confidence of ratings is less high than with tests specifically designed for intermediate or small degradations, such as MUSHRA [23] and BS.1116 [24] (see e.g. [41]). In addition, the choice of stimuli may guide the listeners' attention to specific artifacts or perceptual dimensions, reflected in a certain bias of the rating results [42].

Another concern 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 an over-representation of certain degradation types, but rather with the notion of "signal clarity", as explained in the following [43]: When stereoscopic 3D and 2D video sequences are assessed in the same test, for example using an ACR-type paradigm, 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: In an ACR-test, 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.

To circumvent these methodological problems, image and video quality research has re-adopted the approach of paired-comparison (PC) preference tests in recent years (e.g. [44, 45, 46, 43]). 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 SAMVIQ-tests ("Subjective Assessment Methodology for Video Quality", the video-equivalent to MUSHRA [47]), 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 task, 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, for example [43].

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. [44]). 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 test designs are available that lead to quite efficient tests (e.g. [48]).

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 approaches. 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. As a consequence, in the present paper, a paired-comparison paradigm is adopted for sound quality evaluation tests. The details of the respective test-set-up of Experiment 2 on object-specific sound quality evaluation are described below in Sec. 4.

4. Experiment 1: Localization in Ambisonics

The localization of a synthesized point source was investigated for near-field compensated higher-order Ambisonics at 16 different listening positions. Three different circular loudspeaker arrays each having a diameter of 3 m, but with varying numbers of loudspeakers were applied. The distance between the loudspeakers was 17 cm, 34 cm, and 67 cm. The setup and listening positions are given together with the results in Fig. 1.

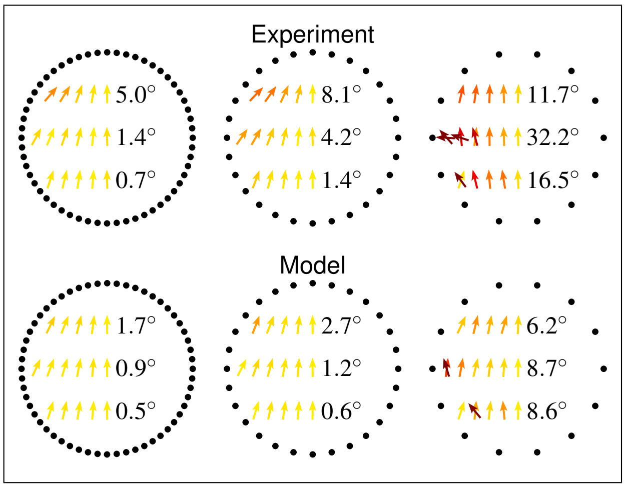

Figure 1. Localization of a synthesized point source in Ambisonics. The top row presents the results from the listening test as mean values over listeners. The bottom row presents predictions of the perceived direction by the binaural model. An arrow is positioned at every applied listening position and points towards the perceived direction of the synthesized source. The color of the arrows indicates the deviation from the desired direction. The more the arrow tends to red, the larger the deviation. Black dots indicate the position of the loudspeakers.

White noise pulses were synthesized as a point source located at $(0, 2.5)$ m. The task for the listeners was to look into the direction from which they perceive the noise, and press a key once their head is correctly oriented. A laser pointer mounted on their heads gave them feedback about their head-orientation. After the key press the next condition started immediately. In the case the listeners perceive more than one source, they were advised to look into the direction of the more pronounced source.

In order to enable a reproducible switching between different loudspeaker arrays and listener positions, the experiment was performed with dynamic binaural synthesis, employing non-individual head-related transfer functions and headphones. The transparency of that method was investigated in more detail in Wierstorf et al. [49].

Twelve listeners participated in the localisation experiment. The top row of Figure 1 summarizes the test results. At every listening position, an arrow indicates the perceived direction of the noise pulse, and its color indicates the deviation from the desired direction. The more the color of the arrow tends to red, the larger is the deviation. Here, optimal localisation is assumed for the targeted direction of the point source. At the right side of every row of listening positions, the value

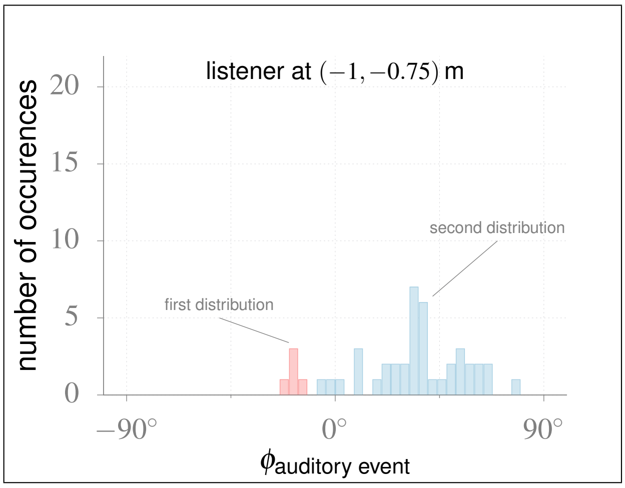

Figure 2. Distribution of localization results shared for all listeners and repetitions per listener for a listener position that yielded more than one perceived position.

of the mean absolute deviation of the perceived direction from the desired one is indicated, averaging over all listening positions included in that row.

At some of the listening positions, the listeners perceived more than one source. This was examined by analyzing the pooled data for all listeners and the five repetitions for every listener and position. The distribution of the directions for a listening position that yielded a perception of two sources is shown in Fig. 2. A perception of two sources only occurred for listening positions to the side of the loudspeaker array and the array with the lowest number of loudspeakers.

In order to model the localization results, a model after Dietz et al. [13] was modified. It is described in detail in Wierstorf et al. [40] and was shown there to predict localization data for Wave Field Synthesis with high accuracy. The challenging task in the present paper was to include a prediction for cases of more than one perceived source. The repetitions for a single listener were simulated by applying different noise sources. With this configuration, the model only showed a single distribution for all listening positions. In a next, step the model was modified so as to turn its head in steps from 0° to 11°, calculating a localisation prediction for the five different noise instances and for all virtual model-head orientations. In this case, the model shows two distributions at some of the listening positions. Figure 1 highlights those positions, indicated by two arrows at one given position that point to different directions.

4.1. Discussion

Obviously, when enabling the model to carry out head movements, the localisation results get closer to the ones obtained with human listeners in localisation tests. However, it must be noted that the model still is not perfectly accurate in predicting human localisation performance. Still, it is able to identify loudspeaker configurations where human listeners heard two instead of one sound source. The exact directions of these two perceived sources cannot yet be accurately predicted by the model. Also, not all positions at which listeners perceived two sources can be identified with the model.

Moreover, so far the implemented head movements do not correspond to those obtained from listeners (see e.g. [37]). Replicating the head movements by the test participants during the tests is a topic of future research. This is easily possible for a large set of the collected localisation data, since head movements have been recorded using an accurate head-tracking system (Polhemus Fastrak).

Another direction for further research consists in the replacement of the binaural model used in experiment 1 by the currently developed TWO!EARS model. One goal here is to further improve the functional agreement between the model and human auditory localisation not only for every-day auditory scene analysis, but also for the case of loudspeaker reproduction. For example, while the model employed in this research needed the head movements to identify critical ambisonics conditions with two perceived sources, human listeners are likely to perceive two sources even without any head-movements. Here, the rich set of bottom-up features currently implemented in TWO!EARS appear to be a promising path towards replicating human-like localisation performance also for the challenging case of spatial audio reproduction.

5. Experiment 2: Scene-based Quality

The perceived quality was investigated using a scene-based paradigm. An excerpt from a music piece consisting of two guitars and a vocal was arranged as a spatial audio scene. The excerpt was taken from the *Blues A* piece, which was recorded at the FH Köln [50] and is available at `http://www.audiogroup.web.fh-koeln.de/anechoic.html`. The guitars were placed to the side of the listener, and the vocals to the front of the listener. For the “ideal reference condition”, the three sources were placed at the corresponding positions indicated in Fig. 3 via head-related transfer functions. In addition, the sources were degraded independently of each other via synthesizing them as focused sources with Wave Field Synthesis, using different loudspeaker setups. The synthesis of focused sources can lead to different impairments of the sound like changes in timbre, its location, or the addition of click-like artifacts [21]. The amount of degradation of a focused source is highly dependent on the position of the listener and the position of the focused source itself. To achieve a similar amount of degradation for the guitar positions and the vocal position, different loudspeaker arrays were applied in both cases, using a linear loudspeaker array for the guitars, and a circular one for

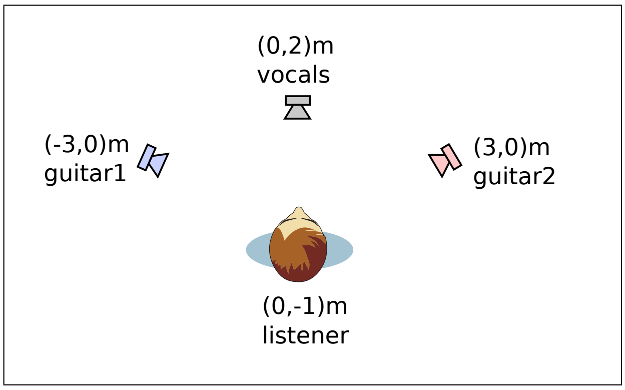

Figure 3. Setup of the quality test.

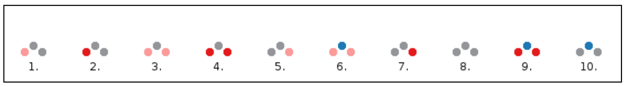

Figure 4. Conditions and quality-rank-order of the scene quality test. Dark and light red represent the two different degradations applied to the guitars, blue represents the degradation of the vocals, and grey indicates the case of no degradation. Condition 8 corresponds to the reference condition.

the vocal. The loudspeaker arrays were simulated by the same binaural synthesis method as applied in the localization experiment, but excluding the dynamic part of switching head-related transfer functions during head movements of the listener. Details of the applied Wave Field Synthesis configurations are described in Dierkes [51]. For the guitars, two different array lengths were applied leading to two different types of degradations for these signals. For the listening test, 10 different conditions[1] were assembled from the different possibilities of combinations. Figure 4 illustrates the chosen conditions.

The choices made for the scene generation obviously do not represent ecologically valid settings, where all degradations stem from the same set-up. This limitation was acceptable for the conducted pilot experiment, since the goal was to investigate whether degradations of different objects in a scene lead to different levels of sound quality.

The task for the listeners was to indicate, in a complete paired comparison test, which of the two presented stimuli was the preferred one in terms of the perceived quality. Twelve listeners participated in the test.

For the analysis of the paired-comparison data, the preference choices of all participants were accumulated in a matrix, to then be transformed to a continuous scale using the Bradley-Terry-model [52, 44]. To this aim, a MATLAB implementation of the model

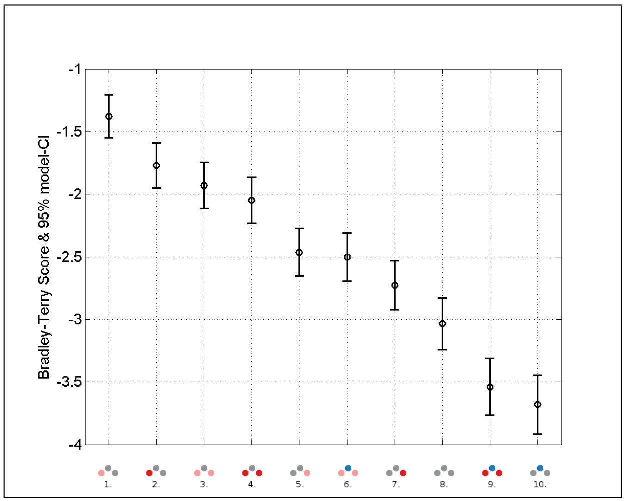

Figure 5. Conditions and ranking of the scene quality test. The paired comparison matrix has been transformed to a continuous Bradley-Terry score. High scores correspond to high quality and vice-versa. See text for details.

was used adopting the calculations previously applied in [48, 43], using the generalized method for model confidence interval calculation proposed by [53].

Here, each entry n_{ij} at position i, j indicates the number of times condition i has been preferred over condition j. Considering that there are N test participants, the respective ratio $p_{ij} = n_{ij}/N$ is used as a likelihood estimation of the probability P_{ij} that condition i is preferred over j. The Bradley-Terry-model relates the preference probability P_{ij} to the quality-distance D_{ij} between the two conditions i and j on a ratio scale. Alternative more general models for PC-data transformation have been proposed in [53].

$$P_{ij} = \tfrac{1}{2}\left[1 + \tanh\left(\tfrac{D_{ij}}{2}\right)\right] \quad (1)$$

$$D_{ij} = \log P_{ij} - \log\left(1 - P_{ij}\right) \quad (2)$$

The resulting scores for the applied ten test conditions illustrated in Fig. 4 are depicted in Fig. 5. The confidence intervals indicate the model confidence for the depicted scale value. On the depicted BT-scale, high values indicate better, and low values indicate worse quality.

5.1. Discussion

It is interesting to note that there are clear quality differences between several conditions. The reference or better source condition #8 clearly is not the preferred one, indicated by the rather low BT-score. Especially cases where the left guitar has been degraded by the focused source settings have achieved the highest preference. Clearly the least preferred are the conditions where the entire scene is rather strongly degraded, and where just the voice is degraded (conditions #9 and #10). It is noteworthy that the reference appears

[1] The stimuli are avialable for listening at `https://github.com/TWOEARS/papers/tree/master/raake2014_forum_acusticum`

to be the third-worst condition. Informal listening reveals a slightly low timbre for the guitars, which is increased with the focused source processing. From the test results and listening to the files it is unclear why condition #6 has been preferred over the reference, too, while the condition with slightly more degraded guitars (#9) or just the voice being degraded (#10) appear to be the only conditions worse than the reference.

The general preference for the degraded guitars may also stem from the fact that this type of instrument is often used with certain effects in popular and rock music, and hence may be preferred when being somewhat distorted. Another interesting finding is that distortion of the left guitar has a more positive effect on preference than distortions of the right guitar. These differences between degradations of the left vs. right guitar can be explained by the fact that the left is being plugged more often than the right one, and hence the general advantage of the distortion on the guitar become more audible.

The results point out that the use of an explicit reference may lead to biased results, or in other words, that there may be clear cases where the reference condition is not preferred over all the other test conditions. Furthermore, experiment 2 clearly indicates that degradations of different objects in an acoustic scene have a distinctly different impact on perceived quality. These findings highlight the adequacy of a scene-based sound quality evaluation paradigm as it is followed by TWO!EARS.

6. Conclusions

An overview of sound quality research has been provided in this paper, followed by the description of two experiments into sound quality evaluation of multichannel loudspeaker reproduction. The research is part of the TWO!EARS project, with its goal to develop a modular test-bed for audio quality modelling. With the two experiments, it was possible to underline the necessity to include active exploration in the domain of sound quality assessment (localisation test, experiment 1), and to address sound quality research in terms of a scene-based assessment paradigm. First results from a respective paired comparison preference test were presented, where the reference condition was perceived as the third worst from the ten test condition. This way, the disadvantages of both ACR-type and MUSHRA-type tests as they were discussed in this paper can be avoided.

Next steps in the sound quality model development within TWO!EARS address a systematic collection of bottom-up features further elucidating the quality-impact due to different spatial audio reproduction configurations. Further, active exploration will be implemented in a more realistic manner, taking head movements by real listeners into consideration.

One of the primary problems related with sound quality evaluation based on world knowledge is the requirement to actually collect that world knowledge, and to train the quality model with it. To this aim, large sound quality test databases are required. As opposed to the TWO!EARS application of auditory scene analysis, where the general performance analysis of the model can partially be made against automatically labelled meta-information, sound quality evaluation requires actual tests with listeners. To complement the tests to be conducted in TWO!EARS, researchers working on similar topics are asked to contact the authors so as to consider collaboration with the TWO!EARS project, for example by exchanging test databases. As indicated earlier, the TWO!EARS project targets open-source developments, so that models or components from other laboratories can easily be evaluated within the TWO!EARS framework. This aspect is considered to be of particular relevance to the field. Here, TWO!EARS follows the spirit of open auditory model exchange first established in the complementary AABBA project [54] and the Auditory Modelling Toolbox developed therein [55]. In TWO!EARS, substantial extensions are underway, starting from a complete object-oriented modeling approach, including a comprehensive software architecture and addressing aspects such as top-down feedback and cognitive integration of scene-specific auditory features.

Acknowledgement

This research has been supported by EU FET grant TWO!EARS, ICT-618075 (see www.twoears.eu). The authors wish to thank Pierre Lebreton, TU Berlin, and Sascha Spors, University of Rostock, for fruitful discussions.

References

[1] A. Raake, J. Blauert: Comprehensive modeling of the formation process of sound-quality. Proc. IEEE QoMEX, Klagenfurt, Austria, 2013.

[2] J. Blauert, D. Kolossa, K. Obermayer, K. Adiloglu: Further challenges – and the road ahead. – In: The technology of binaural listening. J. Blauert (ed.). Springer, 2013, Kap. 18.

[3] C. Schymura, T. Walter, D. Kolossa, N. Ma, G. Brown: Binaural Sound Source LOcalisation using a Bayesian-network-based Blackboard System and Hypothesis-driven Feedback. Proc. Forum Acusticum, 2014.

[4] V. Pulkki: Spatial sound reproduction with directional audio coding. Journal of the Audio Engineering Society **55** (2007) 503–516.

[5] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev: Spatial audio object coding (saoc)-the upcoming mpeg standard on parametric object based audio coding. Audio Engineering Society Convention 124, 2008.

[6] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer et al.: Mpeg spatial audio object coding – the iso/mpeg standard for efficient coding of interactive audio scenes. Journal of the Audio Engineering Society **60** (2012) 655–673.

[7] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, F. Zotter: Spatial sound with loudspeakers and its perception: A review of the current state. Proceedings of the IEEE **101** (2013) 1920–1938.

[8] U. Jekosch: Voice and speech quality perception — assessment and evaluation. Springer, D–Berlin, 2005.

[9] A. Raake, S. Egger: Quality and quality of experience. – In: Quality of Experience. Advanced Concepts, Applications and Methods. S. Möller, A. Raake (eds.). Springer, Berlin–Heidelberg–New York NY, 2014, Kap. 2.

[10] Q. Q. W. Paper: Qualinet white paper on definitions of quality of experience. 1.1 ed. COST Action IC 1003, S. Möller, P. Le Callet and A. Perkis (eds.), Lausanne, CH-Switzerland, 2012.

[11] S. Bech, N. Zacharov: Perceptual audio evaluation. John Wiley & Sons Ltd, UK–Chichester, 2006.

[12] H. Wierstorf, S. Spors, A. Raake: Wahrnehmung künstlich erzeugter schallfelder. German Annual Conference on Acoustics (DAGA), 2014.

[13] M. Dietz, S. Ewert, V. Hohmann: Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Comm. **53** (2011) 592–605.

[14] J. Blauert, U. Jekosch: Concepts behind sound quality: Some basic considerations. Proc. Internoise 2003, 2003, 72–79.

[15] F. Rumsey: Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. J. Audio Eng. Soc. **50** (2002) 651–666.

[16] J. Berg, F. Rumsey: Systematic evaluation of perceived spatial quality. Proc. 24th Conf. Audio Eng. Soc., 2003.

[17] F. Rumsey, S. Zieliński, R. Kassier, S. Bech: On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality. J. Acoust. Soc. Am. **118** (2005) 968–976.

[18] H. Wittek: Perceptual differences between wavefield synthesis and stereophony. Dissertation. University of Surrey, 2007.

[19] H. Wierstorf, M. Geier, A. Raake, S. Spors: Perception of focused sources in wave field synthesis. J. Audio Engineering Soc. **61** (2013) 5–16.

[20] H. Wierstorf, C. Hohnerlein, S. Spors, A. Raake: Coloration in wave field synthesis. Proc. AES 55th International Conference, 2014.

[21] H. Wierstorf, M. Geier, A. Raake, S. Spors: Perception of Focused Sources in Wave Field Synthesis. JAES **61** (2013) 5–16.

[22] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten: PEAQ - the ITU standard for objective measurement of perceived audio quality. J. Audio Eng. Soc. **48** (2000) 3–29.

[23] ITU–R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems. International Telecommunication Union, CH–Geneva, 2003.

[24] ITU–R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. International Telecommunication Union, CH–Geneva, 1997.

[25] ITU–R BS.1284-1: General methods for the subjective assessment of sound quality. International Telecommunication Union, CH–Geneva, 2003.

[26] ITU–T Rec. P.800: Methods for subjective determination of transmission quality. International Telecommunication Union, CH–Geneva, June 1996.

[27] ITU–T Rec. P.862: Perceptual evaluation of speech quality (pesq). International Telecommunication Union, 2001.

[28] ITU–T Rec. P.863: Perceptual objective listening quality assessment (POLQA). International Telecommunication Union, 2011.

[29] J. G. Beerends, C. Sschmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, M. Keyhl: 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** (2013) 385–402.

[30] F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, R. Conetta, S. George, S. Bech, D. Meares: QESTRAL (part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener. 125th Conv. Audio Eng. Soc., 2008.

[31] B. C. J. Moore, C.-T. Tan: Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion. J. Audio Eng. Soc. **52** (2004) 900–14.

[32] M. Brüggen: Sound coloration due to reflections and its auditory and instrumental compensation. Dissertation. Ruhr-Universität Bochum, 2001.

[33] J. Liebetrau, T. Sporer, S. Kämpf, S. Schneider: Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to multichannel audio. Proc. 40th Int. Conf. Spatial Audio, AES, 2010.

[34] K.-M. S. Jeong-Hun Seo, Sang B. Chon, I. Choi: Perceptual objective quality evaluation method for high-quality multichannel audio codecs. J. Audio Eng. Soc. **61** (2013) 535–545.

[35] A. Härmä, M. Park, A. Kohlrausch: Data-driven modeling of the spatial sound experience. Audio Engineering Society Convention 136, 2014.

[36] A. Raake, J. Blauert, J. Braasch, G. Brown, P. Danès, T. Dau, B. Gas, S. Argentieri, A. Kohlrausch, D. Kolossa, N. L. Goff, T. May, K. Obermayer, C. Schymura, T. Walther, H. Wierstorf, F. Winter, S. Spors: Two!ears – integral interactive model of auditory perception and experience. German Annual Conference on Acoustics (DAGA), 2014.

[37] C. Kim, R. Mason, T. Brookes: Head movements made by listeners in experimental and real-life listening activities. J. Audio Eng. Soc. **61** (2013) 425–438.

[38] S. Lepa, E. Ungeheuer, H.-J. Maempel, S. Weinzierl: When the medium is the message: An experimental exploration of medium effects on the emotional expressivity of music dating from different forms of spatialization. 8th Conference of the Media Psychology Division of Deutsche Gesellschaft für Psychologie (DGPs), 2013.

[39] J. H. Michael Schoeffler: About the impact of audio quality on overall listening experience. Proceedings of the Sound and Music Computing Conference (SMC), 2013, 58–53.

[40] H. Wierstorf, A. Raake, S. Spors: Binaural assessment of multi-channel reproduction. – In: The technology of binaural listening. J. Blauert (ed.). Springer, 2013, Kap. 10.

[41] A. Raake, M. Wältermann, U. Wüstenhagen, B. Feiten: How to talk about speech and audio quality with speech and audio people? J. Audio Engineering Soc. (2012) to appear.

[42] S. Zieliński, F. Rumsey, S. Bech: On some biases encountered in modern audio quality listening tests – a review. Journal of the Audio Engineering Society **56** (2008) 427–451.

[43] P. Lebreton, A. Raake, M. Barkowsky, P. L. Callet: Perceptual preference of S3D over 2D for HDTV in dependence of video quality and depth. IVMSP Workshop: 3D Image/Video Technologies and Applications, Seoul, Korea, 2013.

[44] J. C. Handley: Comparative analysis of bradley-terry and thurstone-mosteller model of paired comparisons for image quality assessment. 2001.

[45] A. Benoit, P. L. Callet, P. Campisi, R. Cousseau: Quality assessment of stereoscopic images. IEEE International Conference Image Processing (ICIP), 2008, 1231–1234.

[46] ITU-T Contr. COM 12-C192-E: Comparison of theacr and pc evaluation methods concerning the effects ofvideo resolution and size on visual subjective ratings. International Telecommunication Union, Geneva, 2011.

[47] ITU–R BT.1788: Methodology for the subjective assessment of video quality in multimedia applications. International Telecommunication Union, CH–Geneva, 2007.

[48] J. Li, M. Barkowsky, P. L. Callet: Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment. ICIP, 2012.

[49] H. Wierstorf, S. Spors, A. Raake: Perception and evaluation of sound fields. 59th Open Seminar on Acoustics, 2012, 263–68.

[50] M. Woirgardt, P. Stade, J. Amankwor, B. Bernschütz, J. Arend: Cologne University of Applied Sciences - Anechoic Recordings. 2012.

[51] J. Dierkes: Qualität räumlicher Audiowiedergabe: ist es szenenspezifisch oder objektspezifisch? Tech. Rept. Technische Universität Berlin, 2014.

[52] R. A. Bradley, M. E. Terry: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika (1952) 324–345.

[53] F. Wickelmaier, C. Schmid: A matlab function to estimate choice model parameters from paired-comparison data. Behavior Research Methods, Instruments, and Computers **36** (2004) 29–40.

[54] J. Blauert, J. Braasch, J. Buchholz, H. S. Colburn, U. Jekosch, A. Kohlrausch, J. Mourjopoulos, V. Pulkki, A. Raake: Aural assessment by means of binaural algorithms–the AABBA project. Proc. 2nd Int. Symp. Auditory and Audiological Research–ISAAR'09, 2009, 113–124.

[55] P. Søndergaard, P. Majdak: The auditory-modeling toolbox. – In: The technology of binaural listening. J. Blauert (ed.). Springer, 2013, Kap. 2.