Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

The Technology of Binaural Listening & Understanding: Paper 346

Assessment of audio quality and experience using binaural-hearing models

Alexander Raake[a], **Hagen Wierstorf**[a]

[a] Audiovisual Technology Group, TU Ilmenau, Germany , {alexander.raake, hagen.wierstorf}@tu-ilmenau.de

Abstract:

The paper presents results on spatial audio quality evaluation and approaches for modeling the data using binaural-hearing models, following the interactive and object-related framework developed in the EC-funded FET-Open project TWO!EARS (www.twoears.eu, see also Raake & Blauert QoMEX 2013, Raake et al. Forum Acusticum 2014). Two types of tests and respective modeling approaches are presented: (1) Feature modeling following the well-established approach of spatial and timbral fidelity prediction. To this aim, a feature-specific set of listening tests was conducted for ground-truth data collection. The resulting feature models apply different parts of the bottom-up auditory processing modules from the TWO!EARS framework. (2) Preference modeling. For collecting the underlying data, a listening test series was conducted applying tailored mixes for different musical pieces with different variations of the mixing choices for specific sources in the scene. Using a full paired-comparison test paradigm, preference ratings were collected from listeners. The respective model involves TWO!EARS' scene segregation for object identification and subsequent object-specific feature extraction.

Keywords: active listening, sound quality, Quality of Experience, binaural, model

Assessment of audio quality and experience using binaural-hearing models

1 Introduction

When listeners are presented with audio signals via headphones or loudspeakers, all elements of the end-to-end chain from creation to presentation may play a role for their perception [21]. Since audio technology and respective contents have been specifically created for listeners, the perceived quality is the primary criterion for judging the success of how the technology has been applied. In the respective literature, the terms *sound quality* and *Quality of Experience (QoE)* are typically used (see also our second paper at ICA 2016, [14]). Sound quality explicitly refers to how the influence of the technical system can be perceived. It is sometimes called *Basic Audio Quality*, in line with the terminology used in standards such as MUSHRA, BS.1534 [6]. Sound quality evaluation for loudspeaker-based systems may follow a “spatial and timbral fidelity” paradigm [18]. It relates to findings that, for stereophonic systems, the variance in sound quality tests is explained to 70% by “timbral fidelity” and 30% by “spatial fidelity” [19]. Assessing QoE relates to the audio experience in a more holistic manner, and implies that the listener is not explicitly aware of the fact that the technology is assessed. Examples for attempts to assess QoE for audio systems can be found in [12, 10, 26]. Due to the general difficulty of (direct) QoE assessment, most of the literature from the audio technology domain is on sound quality,

In this paper, we report on our work conducted on sound quality in the TWO!EARS project. The TWO!EARS project aims to develop an active computational model of auditory perception and experience that operates in a multi-modal context (`www.twoears.eu`, FP7, FET-Open). More information on the model concept can be found in [15, 2]. Evaluating sound quality for *spatial audio systems* is one of TWO!EARS two proof-of-concept applications. The other proof is on solving tasks from a search-and-rescue scenario, using dynamic auditory scene analysis.

There are a number of challenges related with spatial audio sound quality and QoE evaluation and the respective model development, for example (see our analysis in another ICA 2016 paper as well, [14]): (1) 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. Also, they may be characterized by differences in certain features without sounding “degraded”. Hence, spatial audio quality can be difficult to assess in tests or by means of models. (2) It is likely that there is no established reference in the minds of listeners when it comes 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 corresponds to loudspeaker-based stereo. For these technologies, dedicated mixing paradigms and listening habits exist, which so far are not commonly available for other spatial audio reproduction techniques. (3) To assess QoE for arbitrary scenes like a human listener can do requires that the prior knowledge of a human person will be built up during the model development, which exceeds what can be achieved for our proof-of-concept.

Because of these difficulties, it was decided to take a pragmatic approach suited for proof-of-concept, namely to address sound quality in terms of the two separate features *coloration* and *localisation*, and to work towards direct sound quality assessment using paired comparison (PC) tests. Using a PC-type preference test paradigm was motivated by recent findings that asking for *video quality* or *Basic Audio Quality* can introduce a bias in the ratings towards timbral or signal clarity features [27, 1, 9]. PC-type preference tests can avoid such biases and addresses simultaneously the two challenges (1) and (2) stated above. In addition, it can be assumed that in a PC-type preference test under laboratory conditions and the correct choice of stimuli, the resulting judgment is close to an assessment of actual QoE, as the listeners are asked to rate which presentation they prefer, instead of rating sound quality.

In Sec. 2, the paper summarizes our previous work on localisation modelling, which has now been integrated into the TWO!EARS model framework[1]. In the second part of Sec. 2, a coloration model is described, adapting adapting the model from [13] to a series of listening tests we have conducted. Our work addressing sound quality is described in Sec. 3. Conclusions and an Outlook on future work are given in Sec. 4.

2 Assessment and prediction of single sound quality attributes

This Section provides an overview of our listening experiments and model development for the two features “coloration” and “localization”. Here, “localization” may better be termed “localization fidelity”, where localization test data is used for developing a model that replicates, as far as possible, the human rating performance.

2.1 Localization

During the last years we have conducted a large number of listening experiments to investigate the spatial fidelity for different sound field synthesis (SFS) systems. The experiments are summarized in [24]. The main focus of the investigation was on the localization and the locatedness of an auditory event synthesized by such systems.

For different spatial audio systems the ability to synthesize a point source placed at a particular position depends on the number of applied loudspeakers and the position of the listener. From a modelling perspective, predicting localization for SFS is a challenging task, 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 prediction, such as interaural time and interaural level differences (ITDs, ILDs). As such, modelling localization for SFS is of high interest also from a perceptual point of view, as the associated auditory features cannot be found in non-technical / natural sound fields.

The long-term 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

[1] http://docs.twoears.eu/1.3/examples/qoe-localisation

that multi-conditional training is a possible way to achieve this [11]. Since version 1.3 of the TWO!EARS model, we achieved the goal of using a common localization framework for auditory scene analysis and localization prediction for SFS. In both cases a deep neural network (DNN) is trained on multi-conditional data. The only restriction for the SFS case is that the used features are restricted to an upper frequency limit of 1400 Hz, whereas in the general case this limit is 8000 Hz.

The model performance has been compared to the results from the implementation presented in [24]. There, the binaural model after [3] was used in combination with a lookup table and some outlier detection to predict the perceived direction. The results of both modeling approaches are very similar, the version using the DNN is on average 1 deg lower in performance. On the other hand it is able to provide the possibility to localize sources in the whole hemisphere by incorporating head rotations, whereas the model presented by Wierstorf [24] works only in the frontal plane.

2.2 Coloration

Besides a limited localization accuracy for virtual sources, the artefacts resulting from the distances between loudspeakers in SFS systems also affect the sound color of the objects and scene at large. For the SFS method of wave field synthesis (WFS) those errors appear only at higher frequencies – for most setups > 1000 Hz – and we could show that the perceptual influence is stronger on the perceived sound color than on the achievable localisation accuracy [25, 24].

Further investigation on the perceived coloration as presented in [25] 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 [8] method in our simulations and reran the listening test on coloration. Compared to the results of the first coloration experiment [25], a lower number of loudspeakers were found to be sufficient to avoid coloration, however still requiring a loudspeaker spacing of 2 cm in a practical setup, see Fig. 1.

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 timbre feature space to another one. This implies a “reference point” (in the experiment labeled as the reference stimulus) to which the listeners compare the timbral perception of another test stimulus.

In WFS, 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 [24]. This simplifies the prediction of coloration, as we focus on spectral auditory features only. As the basis for WFS coloration prediction, the model proposed in [13] was used. In the original paper it was designed to predict the naturalness of different stimuli under comb-filtering conditions. As this was the only factor changed in the original study, it is very likely that their listeners 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 of a reference stimulus: The excitation patterns are calculated using a gammatone filterbank, with subsequent calculation of the standard deviations (SDs) across the frequency channels of the differences between the two excitation patterns. SDs are calculated for the direct differences and for the differences between their 1st order derivatives. The final difference value is the weighted sum of both standard deviations.

The model has two different sets of parameters for speech and noise/music as stimuli. This is set by the user at the moment, at a later stage this could be done by integrating this information from the classifiers available in the TWO!EARS model. As the model requires the excitation pattern of the reference, this has to be known by the model as well. The basic structure of the model uses a blackboard system with different knowledge sources that implement single aspects. The features of the reference signal can hence be stored inside 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 in case we want to add the ability to change the internal reference in a context-dependent manner.

We applied the coloration model to the listening test results we obtained for WFS. 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 feed the exact same stimuli into the model as we used during the listening test. The stimuli and the code for predicting the results is available online.[2] The audio material of the listening test consisted of speech, pink noise and music, with a length of around 9 s.

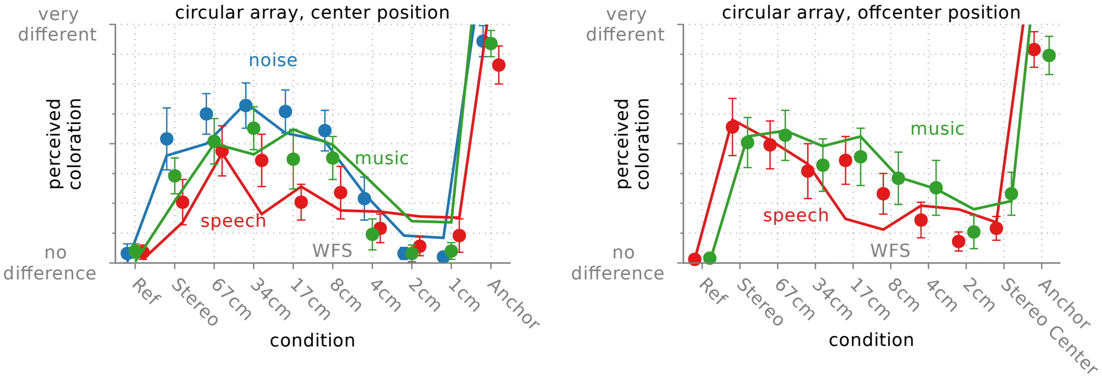

Figure 1: Coloration in WFS for a central and an off-center listening position. The median over 16 listeners together with the confidence interval is shown (points) together with the model predictions (lines). The results for different circular arrays are shown (x-axes ticks: Loudspeaker spacing).

Figure 1 presents exemplary results of the model. 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

[2] See http://docs.twoears.eu/1.3/examples/qoe-coloration/

signal. As expected, the inclusion of the model in the Two!Ears framework is straight-forward, and the model initially developed for other test cases delivers good results also for SFS-type listening stimuli.

2.3 Sound quality and coloration

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 whether two stimuli rated to have the same degree of coloration in comparison with the reference sound are 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 also have the same sound quality rating.

To overcome this problem not only timbral fidelity should be investigated, but listeners should rate the sound quality directly, which is presented in the next section.

3 Assessment and prediction of sound quality

The experiments on localization and coloration presented in the last section were all performed using simple acoustic scenes that consist of one point source. To investigate directly how listeners rate sound quality, a more realistic acoustic scene is required. This can be achieved by using musical pieces, which is a very common usage scenario for a sound reproduction system.

As outlined earlier in this paper and discussed further in [16, 14], the creation and respective recording and mixing process has a strong influence on the reproduced music. Hence, before running tests with realistic musical pieces, some considerations on the mixing process were required. For stereo audio, established processes and respective experience are available for creation, mixing, coding and reproducing audio. For example, level and time differences are used for creating localized “phantom sources”, relying on a channel-based production environment that has achieved a high degree of maturity in practical deployment. This is not the case for sound-field synthesis. For Wave-Field Synthesis, a model-based mixing approach may be used [4]. While different sets of tools are available for this purpose, that plug into commercial systems, there still is a lack of established creation and mixing workflows and rules. Also, stereo mixes cannot directly be transposed to SFS without losing SFS-specific benefits. Similarly, listeners are not yet familiar with the aural experience created by SFS. For both of these reasons, comparison between stereo and SFS-type audio in listening tests is a difficult task. In most studies, simple acoustic scenes have been used, and the aspect of underlying mixes have not been considered.

To address the limitations of previous comparisons between stereo and SFS, specific mixes of four different openly available music pieces were created for 2.1 stereo (X.1: with subwoofer), 5.1 stereo and WFS by the same person [5]. The mixing process was first performed on an independent stereo system where the typical pop-music character was achieved by adjusting

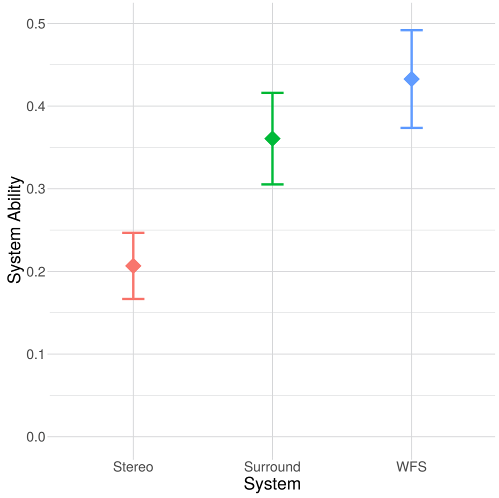

Figure 2: Results of the paired-comparison listening test regarding reproduction system preference. The plots result from an analysis of the data with the Bradley-Terry-Luce model. The dots mark the generated, unity summing system abilities, estimating the ability of each system to be preferred. The bars denote the corresponding 95% confidence intervals. [5]

the level, equalization and compression. In a second step the final mixing on the actual systems took place, whereby a direct switching between the systems was possible. The pieces were reproduced via the same loudspeaker array and assessed in a PC-type preference test by 24 test subjects. Different fixed combinations of mixing–reproduction were used, so no comparison between musical pieces. Based on analyses using the Bradley-Terry-Luce model [23], WFS was found to be preferred over the two other cases, see Fig. 2.

Due to the direct binding of the reproduction systems with the respective mixing process it cannot be excluded that WFS was preferred for the actual mixes and not the reproduction system. For example, in spite of the systematic and well-controlled mixing procedure [5] it could be that the person creating the mix was more motivated during mixing on a new system than he was for the more common stereo and surround. To directly investigate the influence of the mixing process on the sound quality ratings, a new listening test is currently under preparation. Different settings of individual effects such as compression, reverberation, equalization, and placing of foreground instruments will be applied, departing from the baseline of the existing mix for one of the pieces. The resulting modified mixes will be presented to listeners in another PC preference test.

In a next step the results from the PC preference tests should be predicted by the TWO!EARS model, which will be a challenging task. Existing models such as PEAQ [22] or POLQA [7] deliver quality scores using a comparison of the processed audio signal and unprocessed references, after transformation to the perceptual domain (see also 2.2). This is not applicable in our case as there exists no reference signal. Also, sound quality models may relate to the framework of spatial and timbral fidelity [19]. Here, characteristics of the acoustic scene or technical system are transformed into low-level attributes or intermediate quality-related measures. QESTRAL, for example, predicts spatial fidelity from binaural listening cues such as interaural level and time differences [18]. This model is based on the scene-based paradigm of [17], applying foreground-background separation.

The inclusion of a more complex auditory scene analysis / segregation component is one of the pathways how sound quality models applicable to spatial audio can be improved in the future. In the TWO!EARS model, segregation is available (see for example [20]) and will be used to distinguish single instruments or foreground from background. The general framework to predict sound quality will first predict a set of features that correspond to the changes applied to the different mixes and spatial differences between the systems. Those features will then be combined to predict sound quality.

An absolute, no reference rating is a complicated task as the system has to learn an internal reference first, which has proven to be out of reach for the model, also considering the complexity of the required listening tests and stimulus preparation. As stated above, there is no explicit reference for the case of comparing spatial audio systems. On the other hand listeners never rated an isolated stimulus in the experiments, instead they always compared two versions and decided which of those they prefer. The sound quality model will work in a similar way: the goal is not to predict an absolute sound quality value, but to compare two different versions and estimate which is the preferred one. Here, the overall preferred one can be considered as being closest to some kind of internal reference.

4 Conclusions and Outlook

We have summarized the implementation of different aspects of a sound quality model based on the TWO!EARS-framework, and have provided proof-of-concept showing good prediction performance when compared with listening test data for localization and coloration in SFS. Due to its open-source nature, other researchers can build on our results and develop fully functional sound quality models using more comprehensive test datasets than were available in our project. For training or verification purposes, the TWO!EARS test data is available open-source, too (see `http://docs.twoears.eu/1.3/database/`). During the remaining project time from the finalization of this document and the project end, additional modelling steps will be concluded, such as the scene-based prediction of preference for different mixing choices in a WFS-context. The TWO!EARS model is continuously extended during the project, and different software releases have appeared so far, following a reproducible research paradigm. Comparable aspects to be considered in the future are the attention focussing of listeners, and the inclusion of liking of data. We invite other researchers from the field to collaborate with us, by jointly exploiting existing subjective test databases in corresponding model development.

Acknowledgements

This research has been supported by EU FET grant TWO!EARS, ICT-618075 (see `http://www.twoears.eu`). The authors wish to thank Sascha Spors, University of Rostock, and Jens Blauert, Ruhr-University Bochum, for fruitful discussions.

References

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

[2] J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloglu. Further challenges – and the road ahead. In J. Blauert, editor, *The technology of binaural listening*, chapter 18. Springer, Berlin–Heidelberg–New York NY, 2013.

[3] M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *SpeechCom*, 53(5):592–605, may 2011.

[4] M. Geier, J. Ahrens, and S. Spors. Object-based Audio Reproduction and the Audio Scene Description Format. *Organised Sound*, 15(03):219–227, 2010.

[5] C. Hold, H. Wierstorf, and A. Raake. The difference between stereophony and wave field synthesis in the context of popular music. In *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

[6] International Telecommunications Union. *ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems*.

[7] ITU–T Rec. P.863. *Perceptual objective listening quality assessment (POLQA)*. International Telecommunication Union, CH–Geneva, 2011.

[8] Laakso, Valimaki, Karjalainen, and Laine. Spliting The Unit Delay. *IEEE Signal Processing Magazine*, 1(January):30–60, 1996.

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

[10] S. Lepa, E. Ungeheuer, H.-J. Maempel, and 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. In *8th Conference of the Media Psychology Division of Deutsche Gesellschaft für Psychologie (DGPs)*, 2013.

[11] T. May, N. Ma, and G. J. Brown. Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. In *ICASSP*, 2015.

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

[13] B. C. J. Moore and C.-T. Tan. Development and Validation of a Method for Predicting the Perceived Naturalness of Sounds Subjected to Spectral Distortion. *Journal of the Audio Engineering Society*, 52(9):900–14, 2004.

[14] A. Raake. Views on sound quality. In *Proceedings 22nd International Congress on Acoustics (ICA)*, 2016.

[15] A. Raake and J. Blauert. Comprehensive modeling of the formation process of sound-quality. In *Proc. IEEE QoMEX*, Klagenfurt, Austria, 2013.

[16] A. Raake and S. Egger. Quality and quality of experience. In S. Möller and A. Raake, editors, *Quality of Experience. Advanced Concepts, Applications and Methods*. Springer, Berlin–Heidelberg–New York NY, 2014.

[17] F. Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, 50(9):651–666, 2002.

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

[19] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech. 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):968–976, 2005.

[20] C. Schymura, T. Walther, and D. Kolossa. An active machine hearing system for auditory stream segregation. Technical report, Ruhr-University Bochum, 2016.

[21] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter. Spatial sound with loudspeakers and its perception: A review of the current state. *Proceedings of the IEEE*, 101(9):1920–1938, 2013.

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

[23] F. Wickelmaier and C. Schmid. A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior research methods, instruments, & computers*, 36(1):29–40, 2004.

[24] H. Wierstorf. *Perceptual Assessment of Sound Field Synthesis*. PhD thesis, TU Berlin, 2014.

[25] H. Wierstorf, C. Hohnerlein, S. Spors, and A. Raake. Coloration in Wave Field Synthesis. In *AESC55*, pages Paper 5–3, 2014.

[26] A. Wilson and B. Fazenda. Relationship between hedonic preference and audio quality in tests of music production quality. In *Proceedings IEEE 8th International Conference Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016.

[27] N. Zacharov, C. Pike, F. Melchior, and T. Worch. Next generation audio system assessement using the multiple stimulus ideal profile method. In *QoMEX*, 2016.