

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

Deliverable 6.1.1

Scene model framework for auditory scene-analysis



WP 6 *



November 30, 2014

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

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

Work package: 6
Document number: D 6.1.1
Document title: Scene model framework for auditory scene-analysis
Version: 1

Delivery date: 30. November 2014
Actual publication date: 01. December 2014
Dissemination level: Restricted
Nature: Report

Editor(s): Sascha Spors
Author(s): Sascha Spors, Fiete Winter, Guy Brown, Ning Ma, Dorothea Kolossa, Christopher Schymura, Alexander Raake, Hagen Wierstorf, Ivo Trowitzsch
Reviewer(s): Jonas Braasch, Dorothea Kolossa, Bruno Gas, Klaus Obermayer

Contents

| | | |
|----------|---|-----------|
| 1 | Executive summary | 3 |
| 2 | Introduction | 4 |
| 2.1 | Introduction | 4 |
| 2.2 | Description of Scenarios | 5 |
| 3 | Definition and Status of Scenarios | 6 |
| 3.1 | <i>Dynamic Auditory-Scene Analysis</i> Scenario 1 | 7 |
| 3.2 | <i>Dynamic Auditory-Scene Analysis</i> Scenario 2 | 9 |
| 3.3 | <i>Dynamic Auditory-Scene Analysis</i> Scenario 3 | 11 |
| 3.4 | <i>Dynamic Auditory-Scene Analysis</i> Scenario 4 | 13 |
| 3.5 | <i>Dynamic Auditory-Scene Analysis</i> Scenario 5 | 15 |
| 3.6 | <i>Dynamic Auditory-Scene Analysis</i> Scenario 6 | 17 |
| 3.7 | <i>Dynamic Scene Analysis</i> Audio-Visual Scenario 1 | 19 |
| 3.8 | <i>Quality of Experience</i> Scenario 1 | 21 |
| 3.9 | <i>Quality of Experience</i> Scenario 2 | 23 |
| 3.10 | <i>Quality of Experience</i> Scenario 3 | 25 |
| 3.11 | <i>Quality of Experience</i> Scenario 4 | 27 |
| 4 | Conclusions and Outlook | 29 |
| | Bibliography | 31 |

1 Executive summary

The acoustic signals at the ears serve as input for the auditory scene analysis performed by the human auditory system. The same holds for the human visual system where the eyes provide the input. The perceptual model developed in TWO!EARS relies mainly on the auditory sense but also considers the visual sense for multimodal integration.

For the development and evaluation of the TWO!EARS model, a series of audio-visual scenarios has been defined. These scenarios target key challenges in modeling, as well as the proof-of-concept applications considered in the project, namely Dynamic Auditory Scene Analysis (DASA) and Quality of Experience (QoE) assessment. The scenarios are designed with stages of increasing complexity in order to facilitate stepwise development and evaluation. Means of evaluation and benchmarking are defined within the scenarios.

The scenario-based development and evaluation enables an extended and project-specific way of “unit testing”, an approach typically used in software engineering. Alongside modular testing, it ensures that components from different work packages are integrated appropriately. The list of scenarios contained in this document will be extended throughout the coming project phases. The final set of proof-of-concept scenarios available at the end of the project will serve to demonstrate the capabilities of the system and TWO!EARS’ scientific achievements.

2 Introduction

2.1 Introduction

The TWO!EARS model is intended to overcome the limits of current binaural models in various respects. The task of developing such a model is complex and challenging. In this respect it is important to foster stepwise development and evaluation. In the project this is addressed by defining a series of audio-visual scenarios that form the ground for iterative development and evaluation. This process is illustrated in Figure 2.1. In the development phase of a new feature, the capabilities of the model are under the complexity of the associated scenario. The scenario serves here as a template for the generation of audio-visual input to the model. As soon as the model is able to handle the new challenges imposed by the scenario, evaluation plays an important role. Here the scenario provides a well defined framework for the systematic evaluation and documentation of the reached performance. As such, the scenario-based development and evaluation can be considered as a type of “unit testing”, an approach used in software engineering. The scenarios also foster reproducible research, since these allow to reproduce the documented results of a specific scenario. The model will be published together with functionality which easily allows to instantiate a given scenario and to reproduce its published results.

As a proof of concept the TWO!EARS model is applied in two domains. The first application domain is described in Task 6.1. It constitutes *Dynamic Auditory-Scene Analysis* in potentially adverse environments. The second application domain, as described in Task 6.2, considers estimating the *Quality of Experience* of spatial audio reproduction systems. Scenarios for both areas have been negotiated amongst the partners. They address

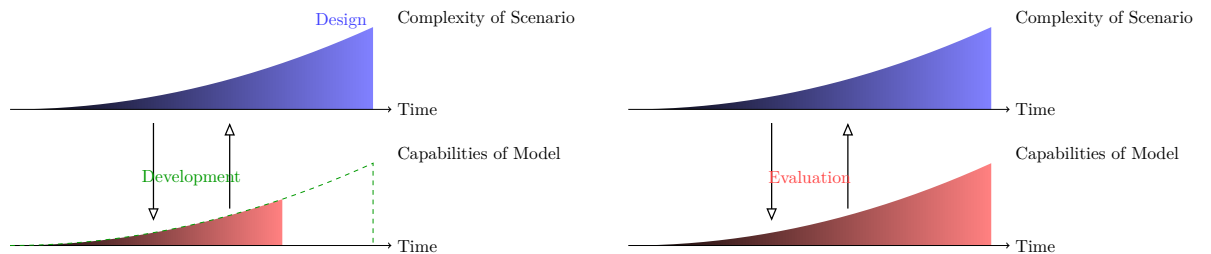


Figure 2.1: Training and evaluation loop using the defined scenarios.

different key challenges and research questions that have to be considered to solve the proof of concepts.

The next Section briefly reviews the information used to describe the scenarios. The scenarios themselves are listed in Section 3.

2.2 Description of Scenarios

The scenarios are described in a unified way. Amongst others, this allows a unified treatment in the simulation framework described in Deliverable 1.1. The description is structured into the following overarching topics

1. description of audio-visual scene,
2. tasks addressed,
3. underlying research questions, and
4. means of evaluation.

The description of the audiovisual scene is composed from information like for instance the audio-visual environment, number and position of objects, position and head-orientation of the listener observing the scene and active acoustic sources. This information forms the basis for synthesizing the ear and eye signals. The parameters are grouped into constant and variable parameters. The latter one constitute the degrees of freedom used for the evaluation.

The tasks which are addressed are also defined in the scenario. They typically constitute basic tasks that form the solution of the potentially complex scenario. The tasks are directly linked to the underlying research questions in context of the scenario. The tasks and research questions are explicitly considered in the evaluation.

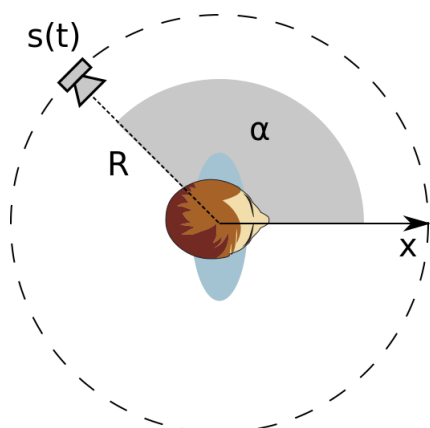
Measures are defined to evaluate in how far the tasks have been fulfilled and research questions have been answered. The documented performance together with the defined scenarios forms the basis to reproduce the model output in the context of reproducible research. This benchmarking includes physical, technical as well as perceptual measures.

3 Definition and Status of Scenarios

The following subsections define an initial set of scenarios negotiated amongst the partners. They are described using the properties and attributes given above. The TWO!EARS software framework allows an easy instantiation of the scenarios as proof of concept and in order to support reproducibility of published evaluation results. The scenarios are ordered by application domain. First scenarios for the *Dynamic Auditory-Scene Analysis* application are considered followed by the ones for *Quality of Experience*. Further scenarios will be defined in the course of the project.

3.1 Dynamic Auditory-Scene Analysis Scenario 1

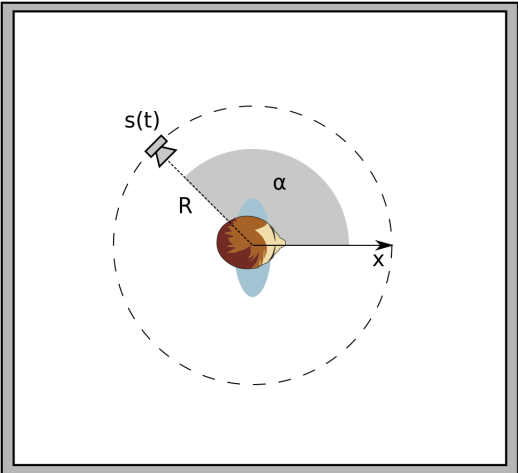
The first *Dynamic Auditory-Scene Analysis* scenario is composed from a single object placed at a fixed distance to the observer in free-space. The task is to locate and/or identify the object by means of the perceived ear signals.

| One source in free-space located at a fixed distance to the listener | |
|--|---|
|  <p>The diagram shows a listener's head in profile, facing right along the x-axis. A source is located at a distance R from the head, at an angle α from the x-axis. A dashed circle represents the source's wavefront. The source signal is labeled $s(t)$. The listener's head is shaded with a blue and brown color gradient.</p> | |
| <p>Constants</p> <ul style="list-style-type: none"> • distance R • initial head orientation • head-related impulse responses (HRIRs) <p>Degrees of freedom</p> <ul style="list-style-type: none"> • incidence direction of source α • source signal $s(t)$ | |
| Tasks | <ul style="list-style-type: none"> • estimate perceived direction • classify source signal |
| Research questions | <ul style="list-style-type: none"> • gain in localization performance due to feedback, e.g. head rotation versus front-back-confusion • gain of source signal classification due to feedback • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |

| | |
|-------------------|--|
| Evaluation | <ul style="list-style-type: none"> • localization accuracy for different source directions α • impact of source signal (e.g. noise, speech, music) • performance of source classification • impact of using different HRIRs for training and localization/classification • changes in the results due to using the Dual-Resonance Non-Linear (DRNL) filterbank model in place of the gammatone filterbank in the Auditory Front-End (AFE) stage |
| Status | <ul style="list-style-type: none"> • a localization knowledge source and a source identification knowledge source have been implemented in WP3 in order to solve the task • a first running version of the complete TWO!EARS model was set up to solve this scenario • the scenario is now a test for the whole model to check if it is still able to solve it after changes we make to it • the evaluation against human data for the different aspects mentioned under evaluation has to be done |

3.2 Dynamic Auditory-Scene Analysis Scenario 2

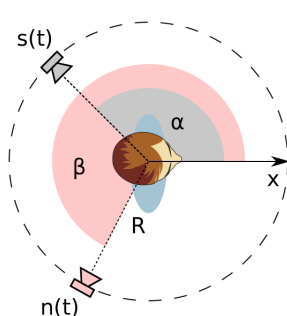
The second scenario extends the first one by considering non free-field conditions. Here a box-shaped reflective environment is considered. The task is to locate and/or identify the object by means of the perceived ear signals.

| One source in a reverberant room located at a fixed distance to the listener | |
|---|--|
|  | <p>Constants</p> <ul style="list-style-type: none"> • dimensions of room • source distance R • initial listener position and head orientation • binaural room impulse responses (BRIRs) <p>Degrees of freedom</p> <ul style="list-style-type: none"> • incidence direction of source α • source signal $s(t)$ • amount of reverberation |
| Tasks | <ul style="list-style-type: none"> • estimate perceived direction • classify source signal |
| Research questions | <ul style="list-style-type: none"> • gain in localization performance due to feedback • gain of source signal classification due to feedback • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |

| | |
|-------------------|---|
| Evaluation | <ul style="list-style-type: none"> • localization accuracy for different source directions α, room geometries, amount of reverberation • impact of source signal (e.g. noise, speech, music) • performance of source classification • changes in the results due to using the DRNL filterbank model in place of gammatone filterbank in the AFE stage |
| Status | <ul style="list-style-type: none"> • the benefit of different head rotation strategies has been investigated using a machine-hearing system for binaural sound localisation in the two reverberant rooms [2]. The performance was improved with head rotation over a no-head-rotation baseline consistently across all the conditions. The best performing head movement among the tested strategies was to rotate the head towards the most likely source direction. • source classification and changing input signals have to be addressed |

3.3 Dynamic Auditory-Scene Analysis Scenario 3

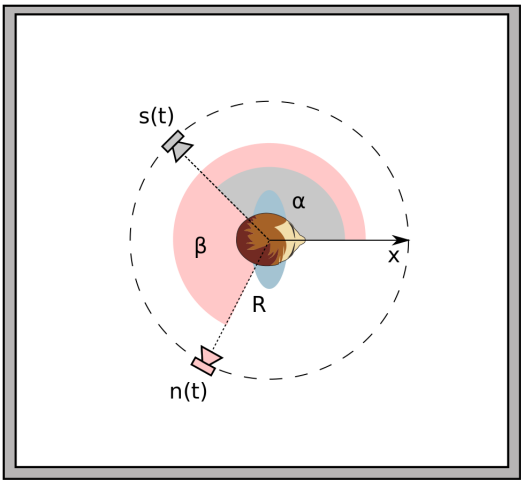
The third scenario considers the distraction by additional objects masking the perception of a desired object. A target and a masker is placed at different positions under free-space conditions. The task is to locate and/or identify the object by means of the perceived ear signals.

| One target and masker in free-space located at a fixed distance to the observer | |
|--|---|
|  | <p>Constants</p> <ul style="list-style-type: none"> • source distance R • initial head orientation • HRIRs • signal of masker $n(t)$ <p>Degrees of freedom</p> <ul style="list-style-type: none"> • incidence direction of source α • incidence direction of masker β • source signal $s(t)$ • signal to masker ratio |
| Tasks | <ul style="list-style-type: none"> • estimate perceived direction of target source • classify target signal • segregate a target signal from background noise |
| Research questions | <ul style="list-style-type: none"> • gain in localization performance due to feedback • which monaural and binaural auditory cues facilitate the segregation of multiple competing sound sources? • gain of source signal classification due to feedback • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |

| | |
|-------------------|--|
| Evaluation | <ul style="list-style-type: none"> • localization performance dependent on the signal to masker ratio • measure how well a target source can be segregated in the presence of background noise, the <i>ideal</i> segregation is referred to as the ideal binary mask (IBM) • changes in the results due to using the DRNL filterbank model in place of gammatone filterbank in the AFE stage |
| Status | <ul style="list-style-type: none"> • in [6] multi-conditional training was used to deal with estimating the azimuth of multiple speech sources. A systematic evaluation revealed that the system was able to generalise well to unseen acoustic conditions, including a different artificial head that was not used for training. • the role of amplitude modulation spectrogram (AMS) features for the task of speech segregation has been investigated in [5, 3, 4]. It was shown that auditory-inspired modulation processing can substantially improve the segregation performance in the presence of stationary and fluctuating interferers. Moreover, a feature normalization stage allowed the segregation system to function over a wide range of signal to noise ratios (SNRs), despite being only trained at low SNRs [4]. |

3.4 Dynamic Auditory-Scene Analysis Scenario 4

The fourth scenario extends the third one to reflective environments. A target and a masker is placed at different positions with fixed distance to the observer in a box-shaped reflective environment. The task is to locate and/or identify the target object by means of the perceived ear signals.

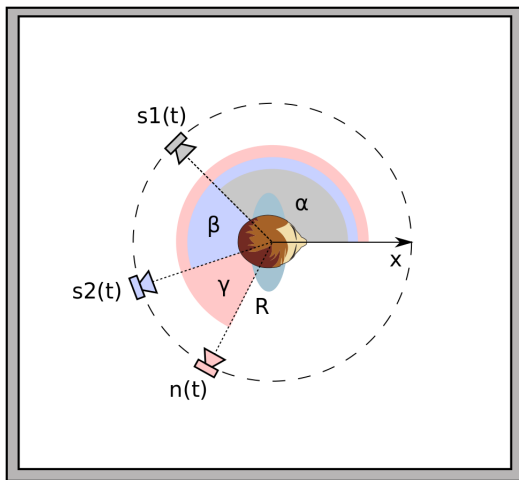
| One target and masker located in a reverberant room at a fixed distance to the observer | |
|---|--|
|  | <p>Constants</p> <ul style="list-style-type: none"> • dimensions of room • source distance R • initial listener location and head orientation • signal of masker $n(t)$ • BRIRs <p>Degrees of freedom</p> <ul style="list-style-type: none"> • incidence direction of source α • incidence direction of masker β • source signal $s(t)$ • signal to masker ratio • amount of reverberation |
| Tasks | <ul style="list-style-type: none"> • estimate perceived direction • classify target signal • segregate a target signal from background noise |
| Research questions | <ul style="list-style-type: none"> • gain in localization performance due to feedback • which monaural and binaural auditory cues facilitate the segregation of multiple competing sound sources? • gain of source signal classification due to feedback • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |

| | |
|-------------------|--|
| Evaluation | <ul style="list-style-type: none"> • how good is the localization dependent on the amount of reverberation measure accuracy of how well a target source can be segregated in the presence of background noise, the <i>ideal</i> segregation is referred to as the IBM • changes in the results due to using the DRNL filterbank model in place of gammatone filterbank in the AFE stage |
| Status | <ul style="list-style-type: none"> • robust localisation of multiple simultaneous speech sources in reverberant environments has been investigated in [6, 2]. It was shown that multi-conditional training using Gaussian white noise can be combined with head rotation to effectively reduce the number of front-back confusions in challenging acoustic scenarios, including multiple competing speakers and reverberation. The system was also able to generalise to unseen acoustic conditions, including a different artificial head that was not used for training. • the source segregation system based on AMS features was evaluated in the presence of room reverberation. It was shown that the feature normalization stage introduced in [5, 4] reduce the sensitivity of the segregation system to room reverberation [5]. |

3.5 Dynamic Auditory-Scene Analysis Scenario 5

The 5th scenario extends the previous one by considering an additional competing object together with a masking object placed in a reflective environment. The task is to locate and/or identify the target object by means of the perceived ear signals.

One target source, one competing source and one masker source located in a reverberant room somewhere on a circle with the listener at the center



Constants

- dimensions of room
- source distance R
- initial listener location and head orientation
- signal of masker $n(t)$
- BRIRs

Degrees of freedom

- incidence direction of target source α and competing source β
- incidence direction of masker γ
- source signals $s(t)$
- signal to masker ratio
- amount of reverberation

Tasks

- estimate perceived direction
- classify source signal

Research questions

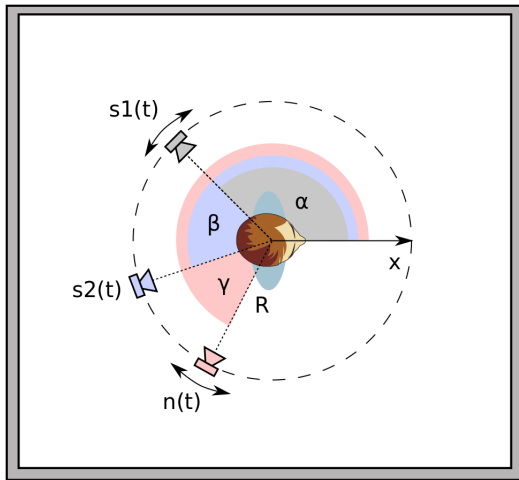
- gain in localization performance due to feedback
- gain of source signal classification due to feedback
- effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance

| | |
|-------------------|---|
| Evaluation | <ul style="list-style-type: none">• how good is the localization dependent on the amount of reverberation• changes in the results due to using the DRNL filterbank model in place of the gammatone filterbank in the AFE stage |
| Status | <ul style="list-style-type: none">• work on this scenario has not started yet |

3.6 Dynamic Auditory-Scene Analysis Scenario 6

So far static objects have been considered, the sixth scenario extends the previous one by considering a moving target and masker object together with a competing object at a static location.

One moving target and masker, and stationary competing source located in a reverberant room at a fixed distance to the observer



Constants

- dimensions of room
- source distance R
- initial listener location and head orientation
- signal of masker $n(t)$
- BRIRs

Degrees of freedom

- incidence direction of source α and competing source β
- incidence direction of masker γ
- source signal $s(t)$
- signal to masker ratio
- amount of reverberation
- speed of target source, masker

Tasks

- estimate perceived direction
- classify target signal

Research questions

- gain in localization performance due to feedback
- gain of source signal classification due to feedback
- effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance

| | |
|-------------------|---|
| Evaluation | <ul style="list-style-type: none">• how good is the localization dependent on the amount of reverberation• changes in the results due to using the DRNL filterbank model in place of gammatone filterbank in the AFE stage |
| Status | <ul style="list-style-type: none">• the identification of time-variant BRIRs has been developed in WP1 in order to accurately capture and synthesize dynamic acoustic environments (see Deliverable 1.1) |

3.7 Dynamic Scene Analysis Audio-Visual Scenario 1

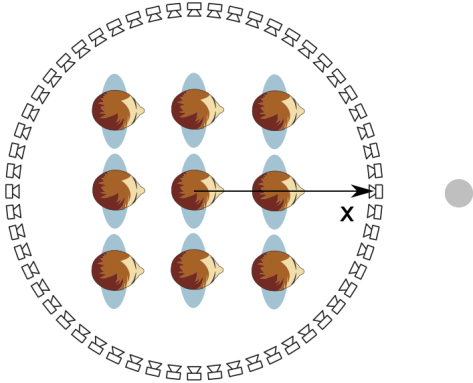
So far only acoustic scenarios have been defined. The Two!EARS model also intends to incorporate other modalities for scene analysis. As a first proof of concept, this scenario constitutes a multimodal extension of Scenario 3. It is constructed from one audio-visual object and one audio masker under free-space conditions.

| One target and masker in free-space located at a fixed distance to the observer | |
|--|--|
| | <p>Constants</p> <ul style="list-style-type: none"> • distance R • initial head orientation • HRIRs • signal of masker $n(t)$ <p>Degrees of freedom</p> <ul style="list-style-type: none"> • source signal $s(t)$ • signal to masker ratio • video signal $v(t)$ • degree of visibility for $v(t)$ • incidence direction of source α • incidence direction of masker β |
| Tasks | <ul style="list-style-type: none"> • estimation of perceived direction • classification of source signal • visual speaker detection/identification • audio-visual speech analysis |
| Research questions | <ul style="list-style-type: none"> • gain in localization performance due to visual feedback • gain of source signal classification due to visual feedback (e.g., audio-visual speaker identification) • gain in the disambiguation of multiple sources due to visual feedback • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |

| | |
|-------------------|--|
| Evaluation | <ul style="list-style-type: none"> • localization accuracy for different source directions α, comparison to human localization performance • impact of cross-modal cue fusion on localization performance • impact of cross-modal cue fusion on classification/identification performance • effect of “mode weighting” to counter adverse environmental conditions (e.g., prefer auditory cues in darkness, prefer visual cues in noisy conditions) • results obtained with the DRNL filterbank model compared to those with gammatone filterbank |
| Status | <ul style="list-style-type: none"> • visual processing stage implemented in the Modular OpenRobots Simulation Engine (MORSE) simulator • face detection enabled for multiple sources • develop integration of visual with auditory cues |

3.8 Quality of Experience Scenario 1

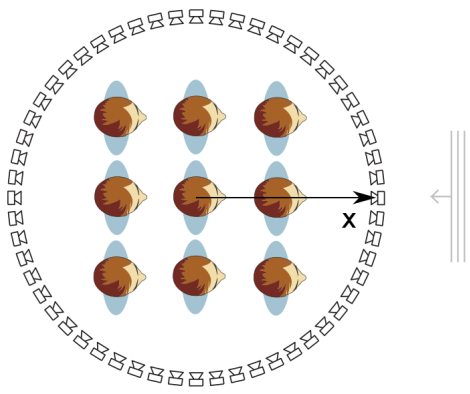
The second application scenario of TWO!EARS focuses on the prediction of *Quality of Experience* for spatial audio reproduction. Consequently several scenarios are defined which cope for this application. The first *Quality of Experience* scenario considers the localization of a virtual point source synthesized by a circular loudspeaker array driven by different sound field synthesis techniques.

| Point source synthesized with different sound field synthesis techniques | |
|--|---|
|  | <p>Constants</p> <ul style="list-style-type: none"> • source type • source position • source signal (white noise burst) • initial head orientation • HRIRs <p>Degrees of freedom</p> <ul style="list-style-type: none"> • listener position • sound field synthesis technique • geometry of loudspeaker array |
| <p>Tasks</p> | <ul style="list-style-type: none"> • estimation of perceived direction • estimation of number of perceived sources |
| <p>Research questions</p> | <ul style="list-style-type: none"> • test best way to identify number of perceived sources • test if model result changes if head movements are allowed • test if the model can automatically find the best head movement • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |

| | |
|-------------------|---|
| Evaluation | <ul style="list-style-type: none"> • compare to results from listening test (see D 1.1) <ul style="list-style-type: none"> – position of auditory event – head orientation performed by listeners • compare to alternative binaural model <ul style="list-style-type: none"> – position of auditory event – number of auditory events • compare the results obtained with the DRNL filterbank model to those with gammatone filterbank |
| Status | <ul style="list-style-type: none"> • localization stage is integrated in Two!EARS model framework • listening test data for evaluation is provided • modeling of the same data is done with the model after Dietz et al. [1] in order to compare the performance of the Two!EARS model • estimation of number of perceived sources has to be developed |

3.9 Quality of Experience Scenario 2

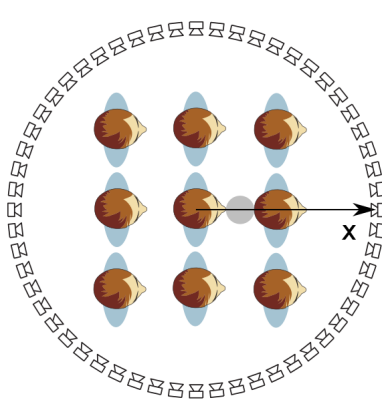
The second *Quality of Experience* scenario is similar to the first one but considers the synthesis of a virtual plane wave instead of a virtual point source.

| Plane wave synthesized with different sound field synthesis methods | |
|---|---|
|  <p>The diagram shows a circular arrangement of 9 small human figures (listeners) seated around a perimeter. A horizontal arrow labeled 'x' points from the center of the array towards the right, indicating the direction of a plane wave. To the right of the array, three vertical lines with an arrow pointing left represent the plane wave source.</p> | |
| <p>Constants</p> <ul style="list-style-type: none"> • source type • source position • source signal (white noise burst) • initial head orientation • HRIR <p>Degrees of freedom</p> <ul style="list-style-type: none"> • listener position • sound field synthesis technique • geometry of loudspeaker array | |
| Tasks | <ul style="list-style-type: none"> • estimation of perceived direction • estimation of number of perceived sources • estimation of source width/locatedness |
| Research questions | <ul style="list-style-type: none"> • what features are needed to model source width/locatedness • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |

| | |
|-------------------|---|
| Evaluation | <ul style="list-style-type: none"> • compare to results from listening test (see D 1.1) <ul style="list-style-type: none"> – position of auditory event – locatedness – head orientation performed by listeners • compare to alternative binaural model <ul style="list-style-type: none"> – position of auditory event – number of auditory events – locatedness • compare the results obtained with the DRNL filterbank model to those with gammatone filterbank |
| Status | <ul style="list-style-type: none"> • localization stage is integrated in Two!EARS model framework • listening test data for evaluation is provided • develop stage that estimates the locatedness |

3.10 Quality of Experience Scenario 3

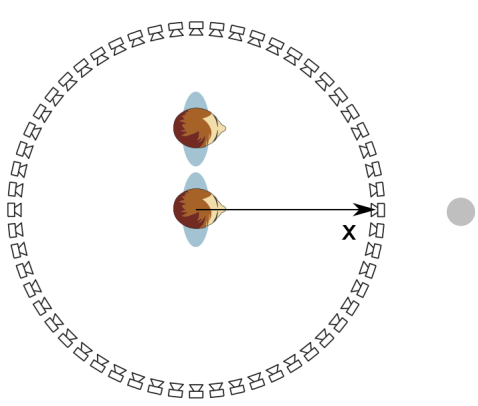
The third *Quality of Experience* scenario is similar to the previous one but considers the synthesis of a focused source instead of a virtual plane wave.

| Focused source synthesized with Wave Field Synthesis | |
|---|---|
|  | |
| <p>Constants</p> <ul style="list-style-type: none"> • source type • source position • source signal (white noise burst) • initial head orientation • sound field synthesis technique • HRIRs <p>Degrees of freedom</p> <ul style="list-style-type: none"> • listener position • geometry of loudspeaker array | |
| Tasks | <ul style="list-style-type: none"> • estimation of perceived direction • estimation of number of perceived sources • estimation of source width/locatedness |
| Research questions | <ul style="list-style-type: none"> • test integration of precedence effect model part • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |

| | |
|-------------------|---|
| Evaluation | <ul style="list-style-type: none"> • compare to results from listening test (see D 1.1) <ul style="list-style-type: none"> – position of auditory event – locatedness – head orientation performed by listeners • compare to alternative binaural model <ul style="list-style-type: none"> – position of auditory event – number of auditory events – locatedness • compare the results obtained with the DRNL filterbank model to those with gammatone filterbank |
| Status | <ul style="list-style-type: none"> • localization stage is integrated in Two!EARS model framework • listening test data for evaluation is provided • develop precedence effect stage of the model |

3.11 Quality of Experience Scenario 4

The fourth *Quality of Experience* scenario is similar to the first one but focuses on the timbral properties of the virtual source instead of localization.

| Coloration of a point source synthesized with Wave Field Synthesis | |
|--|--|
|  <p>The diagram shows a circular arrangement of 32 small speaker icons. Two human figures representing listeners are positioned at the top and bottom of the circle. A central point is marked with a grey dot and an 'x' with an arrow pointing to it, representing the source position.</p> | |
| <p>Constants</p> <ul style="list-style-type: none"> • source type • source position • head orientation • sound field synthesis technique • HRIRs <p>Degrees of freedom</p> <ul style="list-style-type: none"> • listener position • source signal (noise, speech, music) • geometry of loudspeaker array | |
| Tasks | <ul style="list-style-type: none"> • estimation of perceived coloration |
| Research questions | <ul style="list-style-type: none"> • find features and their weighting for predicting coloration • effects of introducing nonlinear peripheral processing model (e.g., basilar membrane nonlinearity) on the performance |
| Evaluation | <ul style="list-style-type: none"> • compare estimated coloration to results from listening test (see D 1.1) • compare the results obtained with the DRNL filterbank model to those with gammatone filterbank |

| | |
|---------------|--|
| Status | <ul style="list-style-type: none">• in a bachelor thesis at TUB different features from the literature on coloration modeling were tested and combined for the prediction of the results• the coloration model has to be integrated into the Two!EARS model framework• run further tests to evaluate the model |
|---------------|--|

4 Conclusions and Outlook

The scenarios listed in this Deliverable serve as a starting point for the development and systematic evaluation of the model. Reproducible research is enabled by providing a simple instantiation of the scenarios and reproduction of published results. For the listed scenarios this process has already begun and will be finished in the second year of the project.

Additional scenarios will be defined during the second year of the project. These include acoustic environments with a more complex structure, like for instance coupled rooms and diffuse noise sources. Additionally multimodal scenarios will be defined to account for audio-visual fusion. The methodology presented in this Chapter serves as template for the definition of new scenarios.

Bibliography

- [1] Dietz, M., Ewert, S. D., and Hohmann, V. (2011), “Auditory model based direction estimation of concurrent speakers from binaural signals,” *53*(5), pp. 592–605. (Cited on page 22)
- [2] Ma, N., May, T., Wierstorf, H., and Brown, G. (2015), “A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions,” in *submitted to the Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (Cited on pages 10 and 14)
- [3] May, T. and Dau, T. (2014), “Requirements for the evaluation of computational speech segregation systems,” *Journal of the Acoustical Society of America* **136**(6), pp. EL398–EL404. (Cited on page 12)
- [4] May, T. and Dau, T. (2014), “Computational speech segregation based on an auditory-inspired modulation analysis,” *Journal of the Acoustical Society of America* **136**(6). (Cited on pages 12 and 14)
- [5] May, T. and Gerkmann, T. (2014), “Generalization of supervised learning for binary mask estimation,” in *International Workshop on Acoustic Signal Enhancement*, Antibes, France. (Cited on pages 12 and 14)
- [6] 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 *submitted to the Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (Cited on pages 12 and 14)