# Two!Ears – Integral interactive model of auditory perception and experience

A. Raake[1], J. Blauert[2], J. Braasch[3], G. Brown[4], P. Danès[5], T. Dau[6], B. Gas[7],
S. Argentieri[7], A. Kohlrausch[8], D. Kolossa[2], N. Le Goff[6], T. May[6], K. Obermayer[1]
C. Schymura[2], T. Walther[2], H. Wierstorf[1], F. Winter [9], S. Spors [9]

[1] *TU Berlin, Germany, alexander.raake@tu-berlin.de*, [2] *Ruhr-University Bochum, Germany*, [3] *Rensselaer Polyt. Inst., USA*
[4] *University of Sheffield, UK*, [5] *Université Toulouse/CNRS, France*, [6] *Technical University Denmark*
[7] *Université Pierre et Marie Curie, Paris, France*, [8] *Technische Universiteit Eindhoven, The Netherlands*
[9] *University of Rostock, Germany*

## Introduction

In the Two!Ears project (`www.twoears.eu`, FP7, FET-Open), we develop an integral, multi-modal, intelligent, active computational model of auditory perception and experience – with two ears and eyes. In its most complete implementation, it will consist of a robot system that can interactively explore its environment based on audio-visual information. The system can serve as a test-bed platform to enable benchmarking of different algorithms for bottom-up signal processing and top-down cognitive processes. The system evaluation targets two applications: (1) Exploratory auditory scene analysis, in terms of a search and rescue task, and (2) Quality of Experience prediction based on interactive exploration of sound fields for evaluating audio reproduction techniques such as wave field synthesis. The system architecture is open and modular, so as to foster progress in perception and experience modelling at large. Specific innovation lies in our interleaved view of bottom-up and top-down processing, our novel expert-system architecture, and in the approach taken for object formation, based on Gestalt principles, meaning assignment, knowledge acquisition and representation, learning, logic-based reasoning and reference-based judgment.

The Two!Ears model builds on a multi-layered architecture with different modules for the bottom-up and top-down processing during interactive exploration (see Fig. 1). For more detailed information on the concepts behind the system can be found in, for example, [1, 2].

## Front-end & acoustic signal processing

The bottom layer represents the physical front-end of the system, and the respective acoustic signal processing.The Two!Ears system will be implemented in a comprehensive software and hardware system. The hardware system will be available in different versions, ranging from static dummy head systems (Kemar) over a head-and-torso-simulator (Kemar-based HATS) with replicas of the pinnae, capable of pan/tilt motion and endowed with cameras for stereoscopic vision, up to a PR2 robot system equipped with the HATS, enabling interactive motion(`http://www.willowgarage.com/pages/pr2/overview`).

The content format for creating interactive acoustic scenes comprises information about the room (acoustic: room impulse response data; geometric: room simulation), position of listener, positions of sources, source types, acoustic contents from different databases such as everyday sounds, speech and music, and semantic annotations describing the scene. For developing the different higher-layer modules, the interconnected software will enable active exploration of virtual or recorded scenes.

## Auditory periphery & pre-segmentation

The layers 2 and 3 address the monaural and binaural subcortical bottom-up processing. The input will be the binaural ear signals from layer 1 representing different scenes with multiple active sources. From this information, primary cues will be extracted:

- Monaural cues: Onsets, offsets, amplitude modulation, periodicity, across-channel synchrony,...
- Binaural cues: Interaural time and level differences (ITDs, ILDs) across frequency bands, interaural coherence (IC), ...

The output of this stage will be a multidimensional auditory representation in terms of "activity maps". These are organized in a topological manner, for example in terms of time, frequency and activity. Based on this multidimensional representation, features for auditory scene analysis are extracted, e.g. features temporally collocated across different spectral bands.

## Meaning assignment

The cognitive components of the system are implemented using a hybrid blackboard/graphical model architecture, with three levels of abstraction. For each layer, there will be a set of expert-systems that carry out specific analysis tasks. Layer 3 addresses pre-segmentation, source separation, visual pattern detection and tracking, and hence comprises low-level experts for pre-segmentation and Gestalt-type analysis. The information is further handled by layer 4, representing event-experts for detecting, classifying and labelling sound events. Meaning is assigned to the auditory events by layer 5, using methods of inference on a graphical model structure. The methods of each layer pass their output information to the underlying multi-level blackboard system. The respective higher-layer experts use this information and related statistical uncertainty data to generate hypotheses. The highest layer 5 reflects cognitive processes, and its exper-

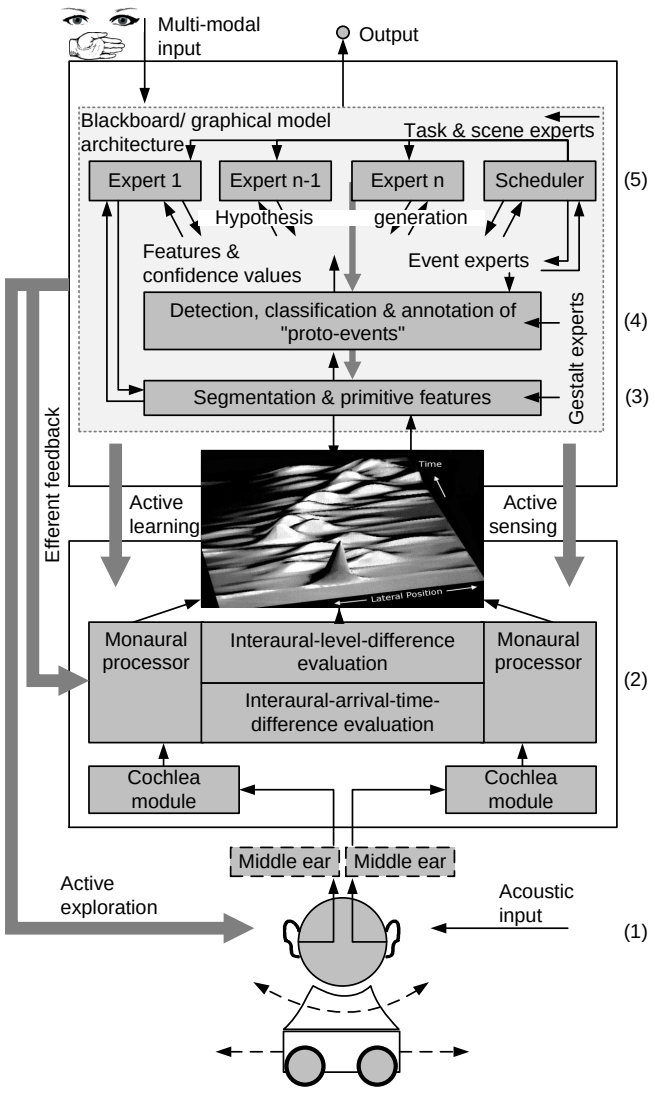


**Figure 1:** Multi-layer TWO!EARS model architecture.

tise has to be based on world knowledge.

The overall functionality can be outlined using an example task, see Fig. 2. Here, it is assumed that the task involves the search for people from acoustic information, that is, the identification and localisation of voice sources. In a top-down-manner, the respective layer 5 experts are activated and launch adaptation calls to lower layers. The event and object experts at layer 4 are thus instructed to focus on speech-like events and adjust their processing correspondingly. To this aim, they may initiate adaptation of the layer 3 processes to provide features for optimal speaker-identification. Likewise, adaptation may address the layers at the auditory periphery level.

## Feedback mechanisms

Implementing top-down feedback is one of the key aspects of the TWO!EARS system. The goal is to use feedback to improve object recognition, auditory grouping, aural-stream segregation, aural-scene analysis, and hence improve the scene understanding, assignment of meaning, attention focusing, and quality judgements. Feedback mechanisms involve both a process that is initiating feedback information, and another process that receives

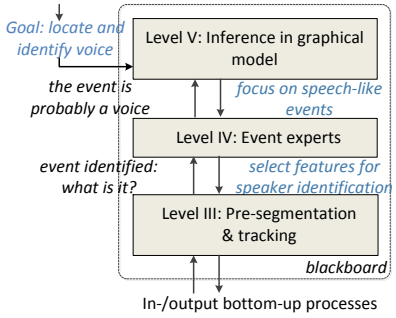


**Figure 2:** Illustration of task-specific scene analysis.

and acts upon it. The following feedback mechanisms will be addressed:

- Turn ears into optimal position (turn-to reflex)
- Advanced exploration of environment by active head-&-torso movements
- Increase signal-to-noise ratio by specific enhancement of spectral & temporal selectivity
- Pay attention to specific signal features, as required by cognitive stage
- Activate specific signal-processing procedures, such as echo cancelling, de-reverberation, precedence-effect preprocessing

## Applications & proof of concept

The overall system shall serve as a testbed for its individual components and alternative modules, for example developed by labs not participating in the project. Two application areas have been selected for the proof-of-concept of the TWO!EARS results, reflecting two dedicated tasks for benchmarking of system variants.

The *dynamic auditory scene analysis* application targets a search- and rescue scenario. Tasks here comprise the dynamic simultaneous localisation and mapping (SLAM) of sources in the scene based on interactive exploration, and achieving higher-level cognitive goals such as sound source or speaker identification, keyword-type speech recognition, and identification of the relevance of sources for the given task.

In the *Quality of Experience* assessment application, the system will interactively evaluate multichannel loudspeaker audio reproduction [1]. It will be investigated, among other aspects, whether active exploration of scenes and individual components lead to better predictions than obtained with traditional approaches, and how internal rather than explicit references can be built into the expert-system to reflect listener expectations.

## References

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

[2] J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloglu, "Further challenges and the road ahead," in *The technology of binaural listening*, J. Blauert, Ed. Springer, 2013, ch. 18.