TWO!EARS http://www.twoears.eu/

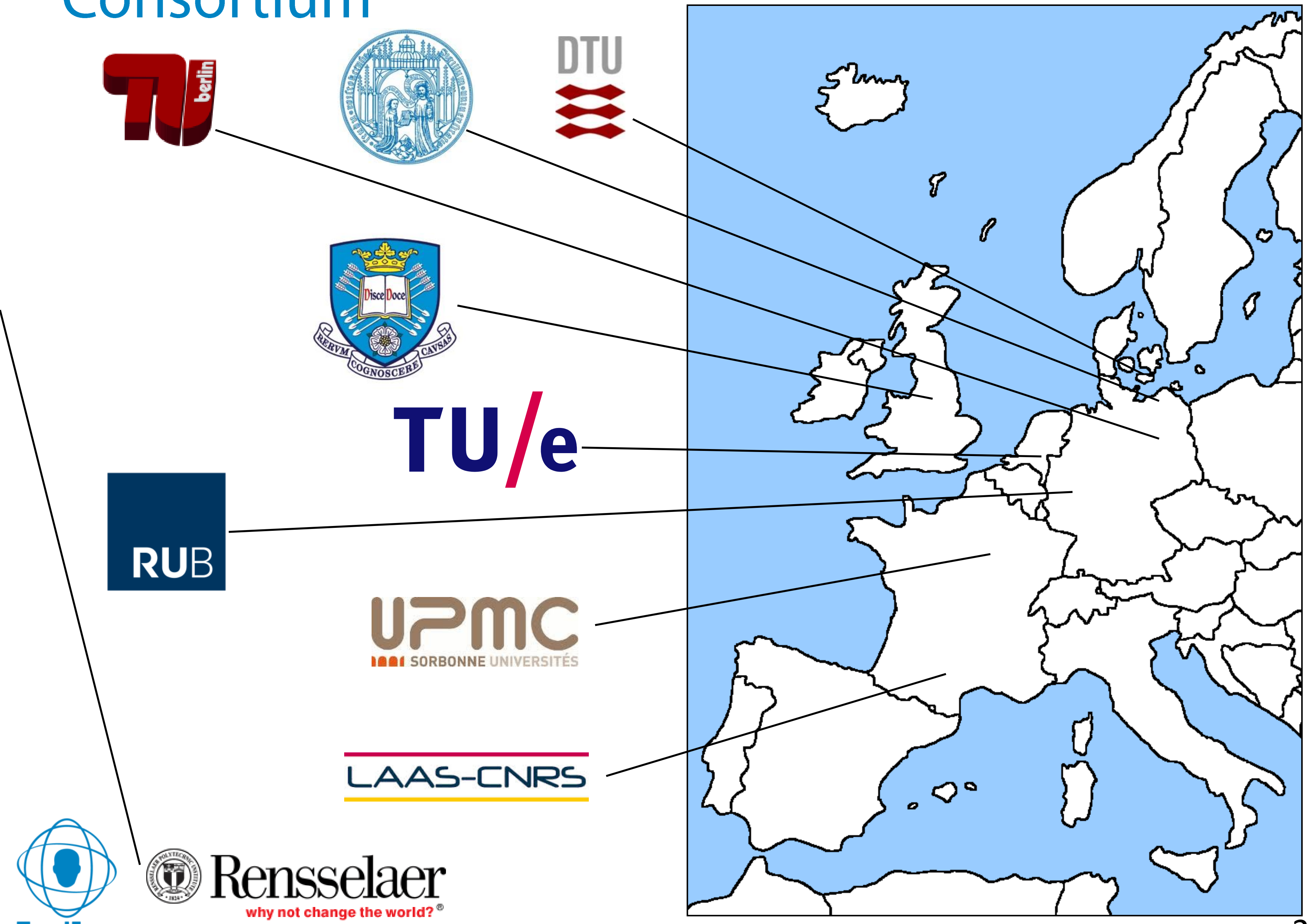
Integral interactive model of auditory perception and experience

Alexander RAAKE
[a]*Assessment of IP-based Applications*
Telekom Innovation Labs (T-Labs)
Technical University Berlin, Germany
www.aipa.tu-berlin.de

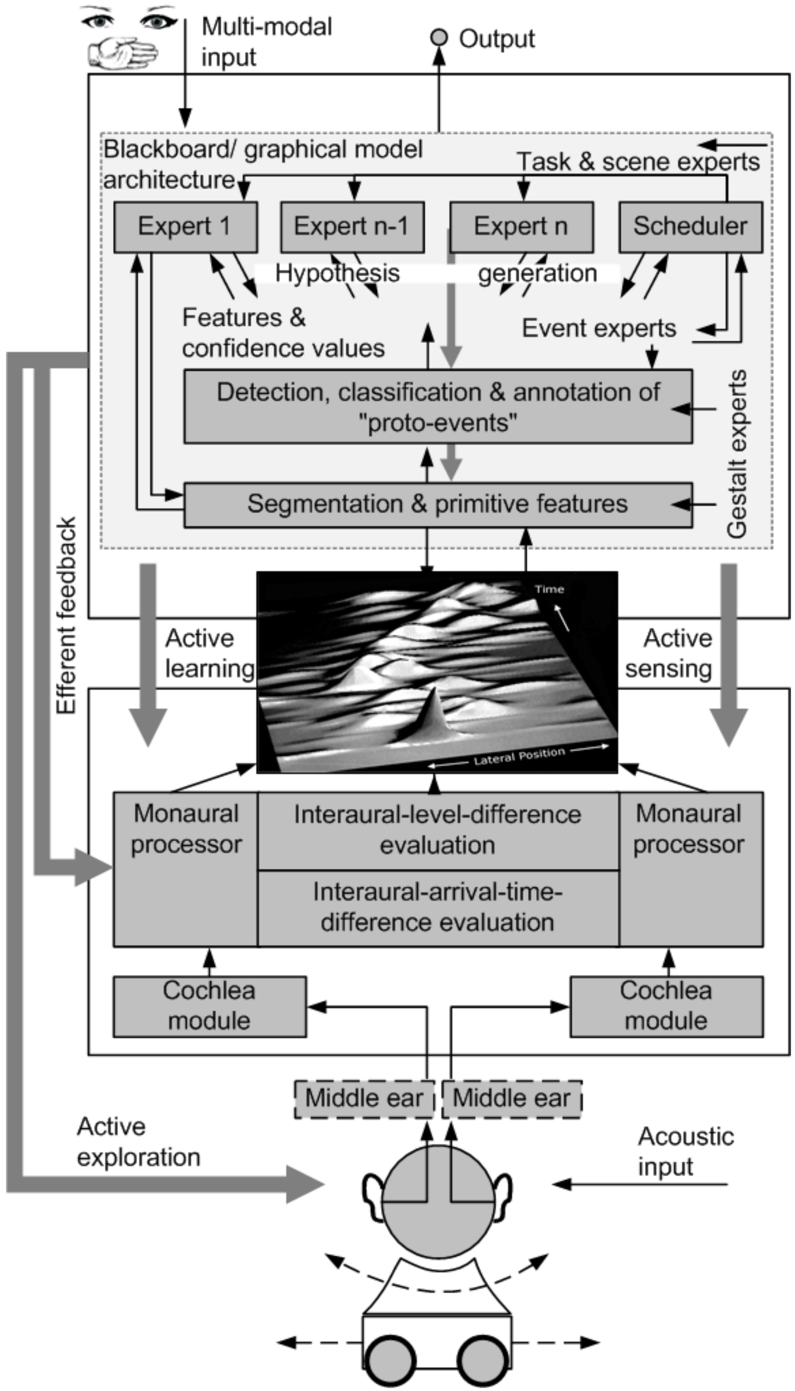
Jens Bauert[b], Jonas Braasch[c], Guy Brown[d], Patrick Danès[e], Torsten Dau[f], Bruno Gas[g], Sylvain Argentieri[g], Armin Kohlrausch[h], Dorothea Kolossa[b], Nicolas Le Goff[f] , Tobias May[f] , Klaus Obermayer[i] , Christopher Schymura[b], Sascha Spors[j] , Thomas Walther[b], Hagen Wierstorf[a]

[b]*Ruhr-University Bochum, Germany;* [c]*Rensselaer Polytechnic Institute, USA*
[d]*University of Sheffield, UK;* [e]*Université Toulouse III Paul Sabatier, France*
[f]*Technical University of Denmark;* [g]*Université Pierre et Marie Curie, France*
[h]*Technische Universiteit Eindhoven, The Netherlands;* [i]*Technische Universität Berlin, Germany*
[k]*Universität Rostock, Germany*

Consortium



TWO!EARS



Level V

Level IV

Level III

Level II

Level I

Level I – Database of scenarios

Settings

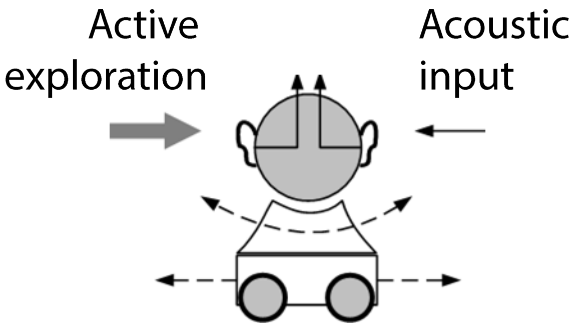
- Defined content formats, shared platform
- Central database of labeled audio-visual scenes & tools for interactive generation
- Head-tracked ear signals incl. translatory movements

Scenarios

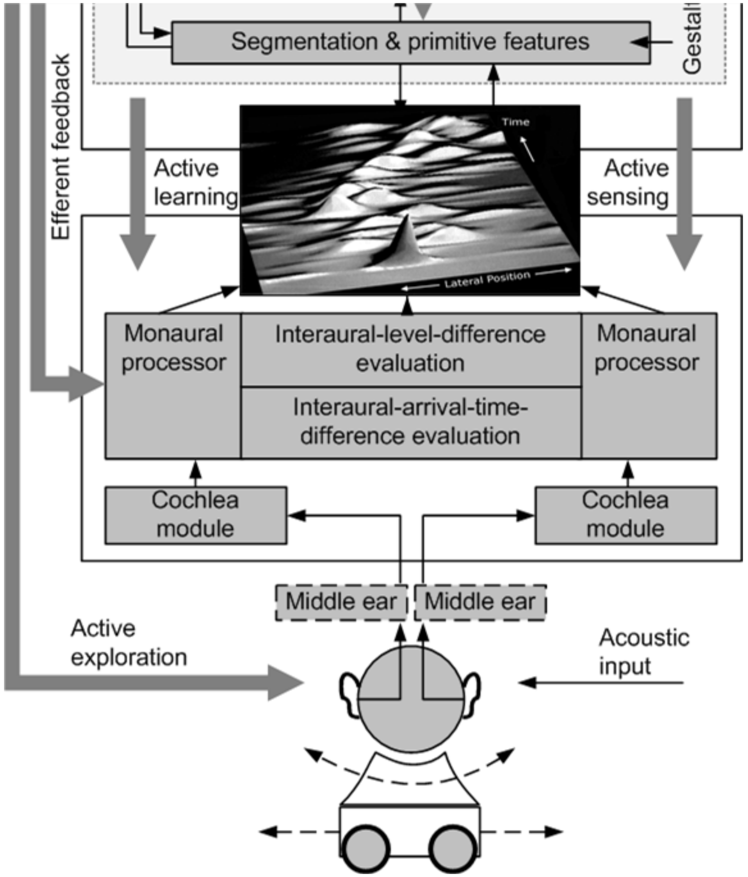
- Natural/synthetic, partly captured with mobile robot platform
- Ear-signals, head-related impulse responses (HRIRs)
- Multichannel recordings, multichannel room-impulse responses
- Still images and video sequences

Dedicated Signal Processing Techniques for dynamic scenes capture

- Advanced techniques for range extrapolation of HRIRs
- Combination of microphone array data with HRIRs
- Capturing of time-variant impulse responses from dynamic sources



Levels II, III



Level III

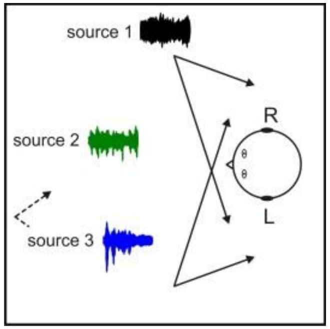
Level II

Level I

Levels II, III

Input level II

- ❑ Binaural ear signals consisting of multiple active sources
- ❑ Feedback from higher stages to adapt bottom-up processing



Extract primary cues

- ❑ Monaural cues: onsets, offsets, amplitude modulation, periodicity, across-channel synchrony
- ❑ Binaural cues: ITDs, ILDs, IC

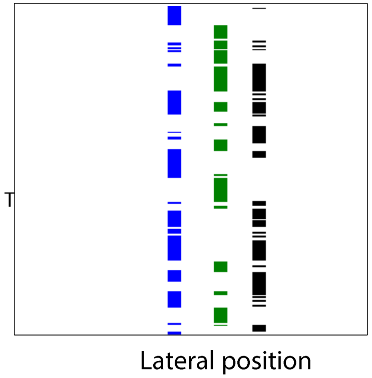
Perform pre-segmentation

- ❑ Determine active sound source positions, detect speech activity (speech segregation), …

Output

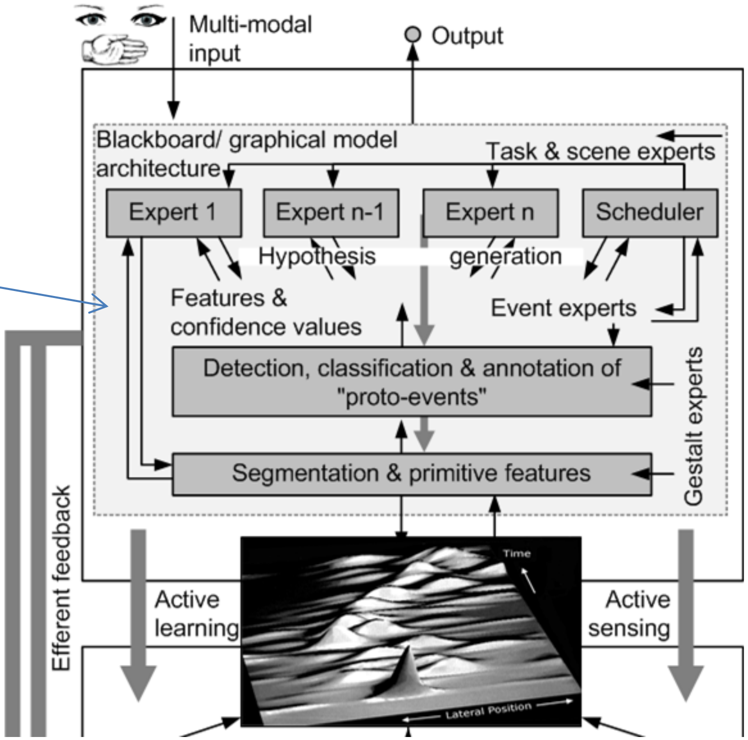
- ❑ Multidimensional auditory representation ("activity maps")
- ❑ Organized in topological manner, e.g. time, frequency, activity
- ❑ Features for auditory-scene analysis (higher levels)
  - – Temporally collocated across different spectral bands
  - – May later be associated with particular objects

Pre-segmented activity map



Levels III-V

- ❑ Blackboard & graphical model architecture
- ❑ Blackboard accessed by different modules
- ❑ Different levels of abstraction

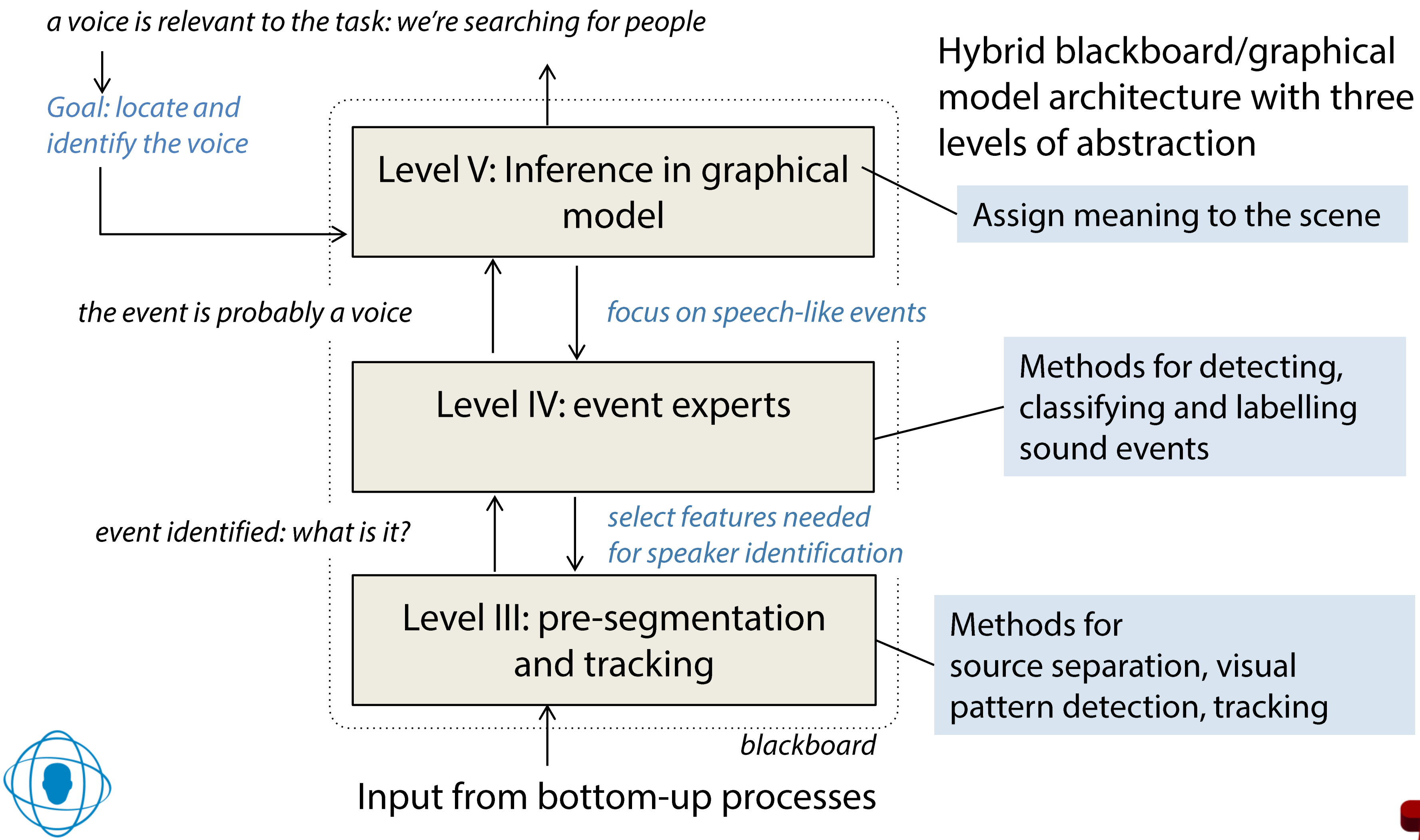


Level V

Level IV

Level III

Level III-V: Example for feature extraction, object formation & meaning assignment


a voice is relevant to the task: we're searching for people

Goal: locate and identify the voice

Level V: Inference in graphical model

the event is probably a voice

focus on speech-like events

Level IV: event experts

event identified: what is it?

select features needed for speaker identification

Level III: pre-segmentation and tracking

blackboard

Input from bottom-up processes

Hybrid blackboard/graphical model architecture with three levels of abstraction

Assign meaning to the scene

Methods for detecting, classifying and labelling sound events

Methods for source separation, visual pattern detection, tracking

Level V

Human cognition

- World-knowledge: Hypothesis generation
- Adaptation & verification processes

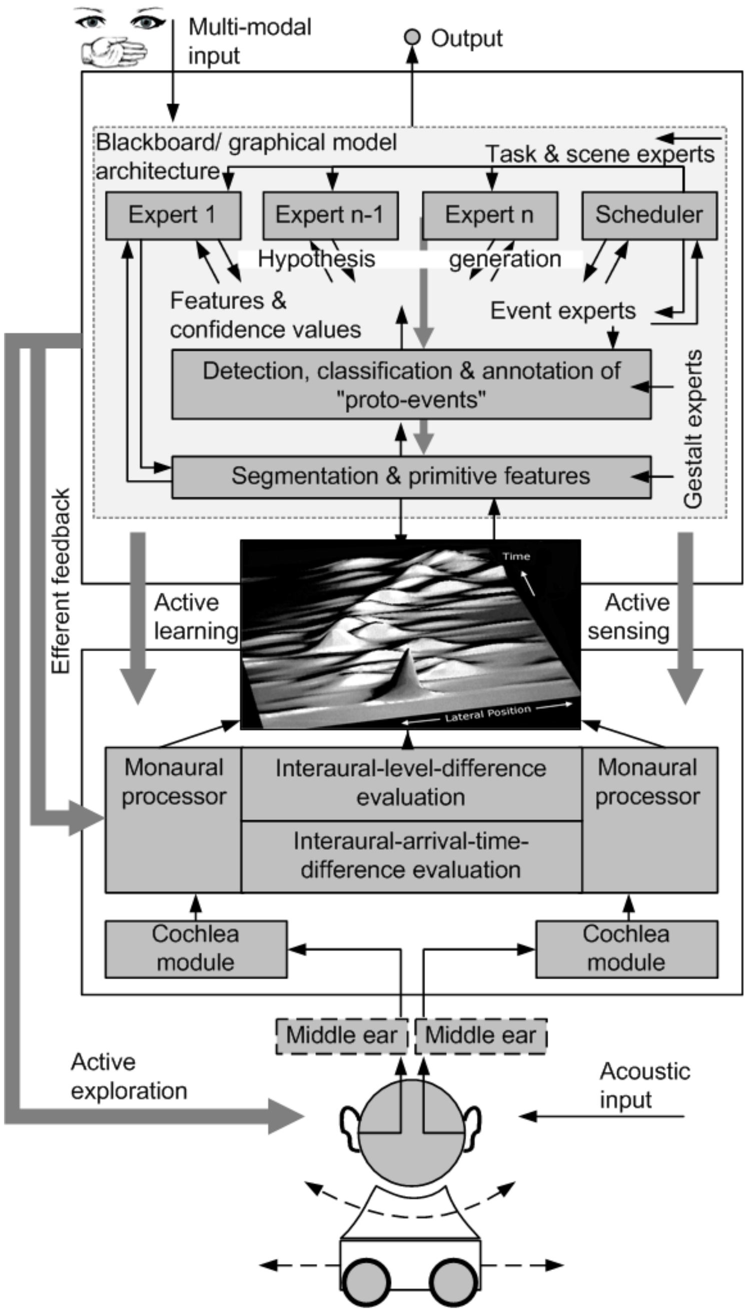
Information accessed by multi-expert system (=software modules)

- Analyze blackboard info (*expertise*)
- Identify whether blackboard information corresponds to available knowledge

Expertise

- Psychoacoustics, object-identification,
- Cross-modal integration, proprioception
- Speech communication,
- Music, sound quality, …

Feedback…



Level V

Level IV

Level III

Level II

Level I

Hardware & software system

Testbed of gradual complexity & versatility

- ❑ Head-&-torso-simulator (HATS)
- ❑ HATS endowed with pan motion and cameras
- ❑ Audio- & visio-auditory head on a PR2 robot

Software architecture

- ❑ Functional & cognitive layers
- ❑ Modular architecture specified with GenoM on the top of the Robot Operating System (ROS) middleware

"Smart" audio-visual sensors

- ❑ Hardcoding of auditory cues with dedicated SoC
- ❑ High-quality cues: Must be embedded, under strong temporal constraints
- ❑ Modular tests under various experimental conditions

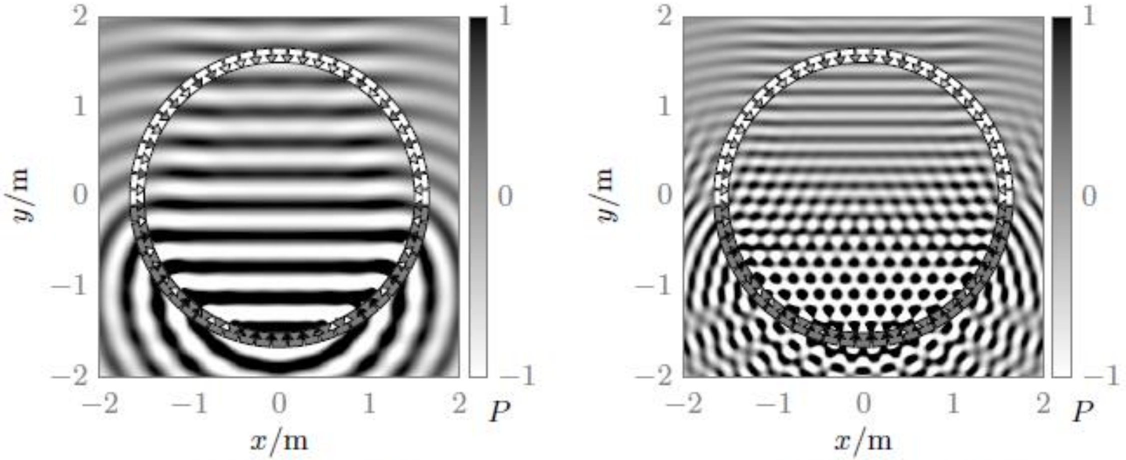
Applications & proof of concept

Dynamic auditory scene analysis – search & rescue

- ❑ Audio SLAM (simultaneous localization and mapping)
- ❑ Speaker identification
- ❑ Keyword-type speech recognition
- ❑ Relevance identification
- ❑ Coarse audio-type identification

Quality of Experience

- ❑ Applied to multi-loudspeaker audio reproduction
- ❑ Active exploration of listening area
- ❑ Internal reference
- ❑ Meaning assignment



(a) $f_{\mathrm{pw}} = 1\,\mathrm{kHz}$

(b) $f_{\mathrm{pw}} = 2\,\mathrm{kHz}$

(Wierstorf et al., this session 10-10:20h)

Summary: Reading the World with Two!Ears

- Functional implementation of active binaural listening & understanding
- Integration of bottom-up & top-down processing
- Computational structure
  - Binaural analysis of acoustic scenes
  - Proprioceptive & visual sensing
  - Active exploration
  - Feedback-based adaptation
  - Cognitive abilities (e.g. attention, source recognition, reasoning, quality)
- Modular test-bed
  - Open software framework
  - Robot for implementation of structure
- Proof-of-concept applications
  - Search-&-Rescue
  - Quality-of-Experience Assessment

Thank you for your attention!
http://www.twoears.eu/