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

Deliverable 6.1.2

Intermediate report on software for analysis of dynamic auditory scenes

WP6 *

November 27, 2015

* 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 packages:	WP6
Document number:	D6.1.2
Document title:	Intermediate report on software for analysis of dynamic auditory scenes
Version:	1

Delivery date:	30th November 2015
Actual publication date:	30th November 2015
Dissemination level:	Restricted
Nature:	Report

Editors:	Guy Brown and Dorothea Kolossa
Author(s):	Dorothea Kolossa, Thomas Walther, Christopher Schymura, Ning Ma, Ivo Trowitzsch, Johannes Mohr, Patrick Danès, Hagen Wierstorf, Tobias May, Guy Brown
Reviewer(s):	Jonas Braasch, Dorothea Kolossa, Bruno Gas, Klaus Obermayer

Contents

1 Executive summary

The Two!Ears project aims to develop an intelligent, active computational model of auditory perception and experience that operates in a multi-modal context. Ultimately, the system must identify the acoustic sources that are present in complex, dynamic environments and ascribe meaning to them. This report has two aims. First, the scenarios that will be used to evaluate the system and drive forward development are described. Secondly, implementation choices that arise from the scenario descriptions are considered.

The scenarios fall into two classes: those used to evaluate the Two!Ears system on dynamic auditory scene analysis tasks, and those concerned with modelling the quality of experience of human listeners. For the former, a number of staged scenarios are proposed which are of increasing complexity. These culminate in a task in which the Two!Ears system – implemented on a mobile robot – must navigate within a multi-room apartment and characterise the sound sources present. The quality-of-experience scenarios focus on the task of predicting human listener's localisation performance and colouration ratings when listening to spatial audio systems.

The second half of this report describes the current blackboard architecture, and the knowledge sources that are likely to be necessary in order to implement the chosen scenarios. The emphasis is on abstract specifications of the knowledge sources, rather than implementation details. The reader is referred to Deliverable D3.4 for details of progress on the implementation of specific knowledge sources.

2 Introduction

2.1 Overview

The Two!Ears project aims to develop an intelligent, active computational model of auditory perception and experience that operates in a multi-modal context. Ultimately, the system must identify the acoustic sources that are present in complex, dynamic environments and ascribe meaning to them. One aim of this report is to formally state the characteristics of those environments, and the tasks that the system must accomplish within them, by setting out a number of *scenarios* that will be used to evaluate the system and drive development forward.

A second aim of this deliverable is to report the current status of the Two!Ears software development. The software will be evaluated within two different types of application according to the work plan of work package 6 (WP6), one of which is concerned with the analysis of dynamic scenes as they are important to understand acoustic and audiovisual scenes, e.g. in search-and-rescue situations. The other utilizes auditory scene analysis to assess the perceptive quality of generated sound fields, with the ultimate aim of comprehensive quality-of-experience prediction.

2.2 Structure of this report

This document first describes the implementation choices that were made regarding the search-and-rescue scenario and regarding the quality-of-experience scenario, cf. Section 3.1 and 3.2, respectively.

The next chapter contains the corresponding software specification, which is composed of two main parts – the specification of the backboard architecture in Section 4.1, and of the knowledge sources in Section 4.2.

The document is concluded by a discussion of the current state of the software and an outlook on the developments in the coming, final year in Chapter 5.

3 Implementation choices

3.1 Search-and-rescue scenario

In the search-and-rescue scenario, the ultimate goal is for the TWO!EARS system, implemented on a mobile robot endowed with acoustic and visual sensors, to locate and characterise specific sound sources in a complex acoustic environment. The system will parse the environment and orient itself according to its intention (e.g., to navigate towards a baby's voice as a matter of priority). In doing so, the system will develop annotated maps of space and knowledge of the system's position within this space. A number of different scenarios will be considered, which vary in complexity.

An initial dynamic auditory scene analysis (DASA) scenario, DASA-1, is only concerned with localizing speakers and with gender identification. A second scenario, DASA-2, will consider keyword recognition. Based on these simple scenarios, two more complex tasks — DASA-3 and DASA-4 — are defined. In these, the system will actually parse and interpret complex environments and aim at identifying possible victims while gaining an understanding of possible dangers. In the final scenario, DASA-4, active exploration will be added, so that the system can move through multiple rooms as necessary until a sufficiently reliable understanding of the scene is attained.

3.1.1 Scenarios

DASA-1: Multi-source speaker localisation and gender recognition

Overview: A female voice is localized in the presence of up to 4 spatially distributed male voice maskers. Two different conditions are considered, one in which the positions of the maskers are known *a priori*, and one in which they are unknown.

Tasks: (i) Find the location of the target voice (ii) determine whether a female voice is present.

Measure of success: (i) root mean square (RMS) error of the target azimuth (ii) Female voice detection performance, quantified by hit rate and false alarm

rate.

DASA-2: Keyword recognition

Overview: Recognition of spoken keywords in the presence of noise and reverberation. Noise conditions include both diffuse background noise and up to 4 interfering sources. An additional database that is used for the evaluation is the CHiME challenge data (Barker *et al.*, 2013), where recordings of domestic noise in a living room (e.g. vacuum cleaners, children playing, music) are superimposed on binaural speech recordings.

Tasks: Identify the keyword that was spoken.

Measure of success: Word error rate.

DASA-3: Localisation and characterisation of sources in a single room

Overview: This is a simplified version of the search-and-rescue scenario in which particular sound sources of interest ('victims') are situated in a single room. Detection of victims involves localisation, male/female or young children's voice detection and keyword recognition in the presence of acoustic clutter. Feedback methods are employed to disambiguate the input data, with limited movement of the robot platform (e.g., head rotation); for more information please see deliverable D4.2, which gives a specification of feedback loops and reports on implementation progress.

For this scenario, it is therefore necessary to detect persons visually, localise sound sources, segregate sound sources, and classify sound sources, including keyword recognition. Identification of sound source characteristics that are relevant to the search-and-rescue scenario (e.g., the detection of distressed speech) is also required.

Tasks: (i) Identify the location of sound sources and orient the head of the robot towards each of them in turn (ii) identify the gender of voices in the scene (iii) source classification (iv) identification of spoken keywords (v) detection of distressed speech.

Measure of success: (i) RMS error of the target azimuth (ii) gender recognition rate (iii) source classification rate (iv) word error rate (v) ability to detect distressed speech (hit rate and false alarm rate).

DASA-4: Localisation and characterisation of sources in a multi-room apartment

Overview: This scenario concerns a search-and-rescue task in a complex multi-source environment. The environment consists of three coupled rooms, in which the robot moves freely in order to detect sources of interest ('victims') and sources to avoid ('hazards'). Detection of victims involves male/female or young children's voice detection and keyword recognition in acoustic clutter. Detection of stressed voices will also be addressed. Feedback methods are employed to disambiguate the input data (e.g. by moving the head or orienting the entire robot with respect to a specific sound source) and to move through the rooms as necessary until a sufficiently reliable environmental map has been created.

For this scenario, it is again necessary to detect persons visually, localize sound sources, segregate sound sources, and classify sound sources, including keyword recognition. Additionally, it is necessary to decide on the optimal movement of the system that is most likely to improve the environment map of the most salient sources.

Tasks: (i) Identify the location of sound sources and orient the whole robot towards each of them in turn (ii) identify the gender of voices in the scene (iii) source classification, including discrimination of victims and hazards (iv) identification of spoken keywords (v) detection of distressed speech (vi) navigation through the environment.

Measure of success: (i) RMS error of the target azimuth (ii) gender recognition rate (iii) source classification rate (iv) word error rate (v) ability to detect distressed speech (hit rate and false alarm rate) (vi) time taken to navigate through the environment, identifying the location of each target source in the process.

3.1.2 Implementation

Under the control of the scheduler in the blackboard system[1], the appropriate knowledge sources for each of the respective tasks need to be called in an appropriate order, which may be determined either by a task-dependent recipe, or by calling knowledge sources in response to the current blackboard state.

1 `http://twoears.aipa.tu-berlin.de/doc/1.0/blackboard/knowledge-sources/`

DASA-1

For the DASA-1 scenario, knowledge sources will be required for *source segregation*, *source localization* and *gender recognition*. We consider two sub-scenarios, which differ according to whether prior information is available or not.

In sub-scenario 1, where no prior information is available, the task will be carried out by an inital source segregation, followed by localization and gender recognition for all source directions.

In sub-scenario 2, training data is available to initialize the localization estimates of the interferers. After this initialization of localizations has been carried out, the source segregation is carried out including the prior localization information. As in the first case, the final step involves gender recognition for all segregated sound sources.

In both cases, the output of the system is the localization that is attributed to the recognized female voice – if a female voice has been identified – and an empty element, in case no female voice has been found among the segregated outputs.

DASA-2

For the DASA-2 scenario, knowledge sources will be required for *source segregation*, *source identification* and *keyword recognition*.

After an initial source segregation, the source identification determines, for each of the sources, whether it is a speech signal. For all speech signals that are detected, the keyword recognition knowledge source is activated.

As an optional, improved version of this scenario, the segregation can be refined by including information about the source identity – from the source identity KS – and, as applicable, the phonetic state, from the keyword recognition KS. Both could be included e.g. as priors on the acoustic features. The refined segregation hypothesis can be used in a second pass of source identification and keyword recognition, and this iteration can be carried out multiple times, e.g. until convergence.

DASA-3

For the DASA-3 scenario, knowledge sources will be required for *visual person detection*, *planning head rotations*, *source segregation*, *source identification*, with provisions for *identification of distressed speech* and *identification of alarm sounds*, *gender recognition*, and *keyword recognition*.

In this scenario, the scheduling is determined not from a simple recipe, but dynamically, corresponding to the status of the blackboard. Various control strategies for the scheduler are considered and compared. One interesting strategy is described in the following:

Phase 1: After an initial source segregation, the source identification determines, for each of the sources, whether it is a speech signal. For all speech signals that are detected, the keyword recognition knowledge source and the distress detection knowledge source is activated.

Phase 2: Since the DASA-3 system follows multiple goals (RMS target localization, gender recognition rate, source classification accuracy, gender recognition accuracy, ability to detect distressed speech), in Phase 2, the initial hypothesis of Phase 1 is successively refined. This refinement is guided by the confidence of the system regarding the fulfillment of each of its goals. For example, a low confidence in the target localization can lead to activating the head rotation Knowledge Source (KS), and a low confidence in the segregation output can lead to an iterative refinement of source segregation by employing source models. Phase 2 continues, until all variables have been determined with sufficient accuracy.

DASA-4

For the DASA-4 scenario, knowledge sources will be required for *visual person detection*, *planning head rotations*, *planning robot movements*, *source segregation*, *source identification*, with provisions for *identification of distressed speech* and *identification of alarm sounds*, *gender recognition*, and *keyword recognition*.

Similarly to DASA-3, scheduling is determined not from a simple recipe, but dynamically, corresponding to the status of the blackboard. Various control strategies for the scheduler are considered and compared. The strategy described above is again an example of a possible scheduler control, with the difference that inaccurate source identification or localization can now also be countered by the additional robot movement planning KS.

3.2 Quality of Experience

The final goal of the Quality of Experience (QoE) application of the Two!Ears model is to predict the perceived quality in different spatial audio scenarios (Raake *et al.*, 2014). It is planned to have model predictions based on a given explicit reference and to have model predictions without an explicit external reference. In the latter case, an internal reference has to be learned, which might change depending on the experience of the model. Another goal is to make use of the opportunities that

dynamic scene exploration adds to quality perception in spatial audio systems and to include context information into the modelling process. The quality-of-experience application relies heavily on the availability of results from listening experiments as most of the questions we are targeting to model have not been considered in the literature yet. As at the current time not all listening tests are finished (in accordance with the workplan), the list of presented quality-of-experience scenarios will contain two pure modeling scenarios, where all the listening tests have been done already. In addition, two scenarios are presented, where the listening experiments are in the preparation phase at the moment.

3.2.1 Scenarios

QoE-1: Predict localisation accuracy in spatial audio systems

Overview: The Two!Ears database already includes several results on localisation accuracy for different sound field synthesis systems[2]. The goal is to predict the localisation by the model.

Tasks: (i) Find the direction of the synthesized source. (ii) Detect, if more than one source is present.

Measure of success: Compare deviation from listening tests results.

QoE-2: Predict coloration in spatial audio systems

Overview: The Two!Ears database already includes listening test results on coloration in wave field synthesis[3] and will be extended by more results for wave field synthesis, as well as for local sound field synthesis techniques with D 1.2. The goal is to predict the coloration ratings with the model.

Tasks: (i) Predict coloration ratings in comparison to a given reference for different spatial audio systems and audio source materials.

Measure of success: Compare deviation from listening test results.

2 http://twoears.aipa.tu-berlin.de/doc/latest/database/experiments/#localisation-of-different-source-types-in-sound-field-synthesis

3 http://twoears.aipa.tu-berlin.de/doc/latest/database/experiments/#coloration-of-a-point-source-in-wave-field-synthesis

QoE-3: Finding the sweet-spot in 5.0 surround systems

Overview: This scenario investigates the dynamic exploration of the listening area in 5.0 surround setups. The listeners should find a point at which they prefer the actual reproduction. Furthermore, the influence of the visual modality on the task is considered, as the listener could get visual feedback on his or her current position and movement in some cases. In order to achieve this, the corresponding listening test will be performed with binaural synthesis. Nine different listening positions in a 5.0 surround setup will be simulated. The listeners then have to rate at which simulated position they prefer to listen to the presented recording. Due to the use of binaural simulation there will be no real loudspeaker setup and the listener will always sit at the same real position in the listening room. Some listeners will get visual feedback on the locations of the simulated positions via a graphical interface. Others will get no feedback about the locations.

The modeling of the results will include both dynamic scene exploration and context information in the form of the visual feedback on the position.

Tasks: Find the position where the listener would prefer to listen to the presented audio.

Measure of success: Deviation from listening test results.

QoE-4: Comparison of spatial audio mixes for wave field synthesis and 5.1 surround stereophony

Overview: In order to be able to compare the achievable Quality of Experience with different spatial audio systems it is important to investigate not only single point sources and ask for localisation accuracy and coloration, but have complex scenes that make use of the advantages provided by the different systems. Here, we create mixes of popular music for 5.1 surround and wave field synthesis systems to compare them directly in a listening test. The goal is to find the underlying perceptual attributes that are used during the comparison by listeners and to model them.

Tasks: Compare surround and wave field synthesis mixes of the same piece of music and say which one a listener would prefer.

Measure of success: Deviation from listening test results.

3.2.2 Implementation

In this section, we will discuss the implementation details only for the first two quality-of-experience scenarios as the details for the second two scenarios can not

be fixed before the results from the listening tests are available (some of which are scheduled later in the work plan).

QoE-1

For the QoE-1 scenario, knowledge sources will be required for localisation and estimation of number of sources.

For the localisation knowledge source, first tests have shown that the ability to use head rotations will help to prevent front-back confusions (Ma *et al.*, 2015). In addition, it was observed that the current implementation of the localisation knowledge source, which is only trained under free field conditions, will not be sufficient to predict the data correctly. Instead, we will try to use a localisation knowledge source that is trained under more natural conditions. We have already used multi-conditional training to get good localisation results in rooms (May *et al.*, 2015), which will be tested here as well.

In addition, a knowledge source for estimating the number of sound sources has to be implemented. This is necessary in order to predict some of the results on near-field compensated higher order ambisonics, as the perceived sound source can split into two at some listening positions.

QoE-2

For the QoE-2 scenario a coloration knowledge source will be required, which is able to learn a reference and to judge test conditions on the their deviation in timbre regarding this reference.

The estimation of deviation in timbre is based on the model of Moore and Tan (2004) which is able to predict the naturalness of sounds that have different comb-filter-like patterns applied to their frequency response. It appears that their naturalness ratings are highly correlated to coloration judgments, as the only changes they had in their stimuli was a change in timbre. Furthermore, the model is appropriate to judge coloration in sound field synthesis as the physical changes of the stimuli are very similar to applying comb-filter-like filters to the frequency response.

Learning the reference is done by storing its extracted auditory features, as needed for the coloration model, in the coloration knowledge source.

4 Software specification

This chapter describes the knowledge sources which will most likely be required in order to satisfy the requirements of the scenarios presented in the previous chapter. The focus here is on abstract specifications, rather than implementation details; progress on implementation of many of these knowledge sources is reported in Deliverable D3.4.

4.1 Blackboard architecture

The current structure and function of the blackboard system is based on the architectural considerations that were presented in Deliverable D3.2. It is targeted as the front-end for a great variety of applications, providing an architecture that integrates experience formation and active behaviour from a set of individual functional modules. These modules can work on different levels of abstraction, independently from each other or in collaboration, in a bottom-up or top-down manner. A key feature of this system is its ability to evolve, so that easy modification, exchange and/or extension of modules can be achieved within a scalable architecture. The current implementation of the blackboard system is based on three main components:

Blackboard The blackboard holds the central data repository of the platform. It not only stores current data, but keeps track of the history of this data in order to enable working on time series data.

Knowledge Sources (KSs) are modules that define their own functionality, to be executed in the organised frame of the system. They define themselves, which data they need for execution and which data they produce. The blackboard system provides the tools for requesting and storing this data, but does not care about the actual contents, while the KSs do not need to care about where and how data is stored.

Scheduler The scheduler is the component of the blackboard system that actually executes the KSs – but first, it schedules them, that is, it decides the order in which KSs get executed. This order is rescheduled after every execution of a KS, since the conditions determining the order may have changed, or new KSs may be waiting for execution that are more urgent.

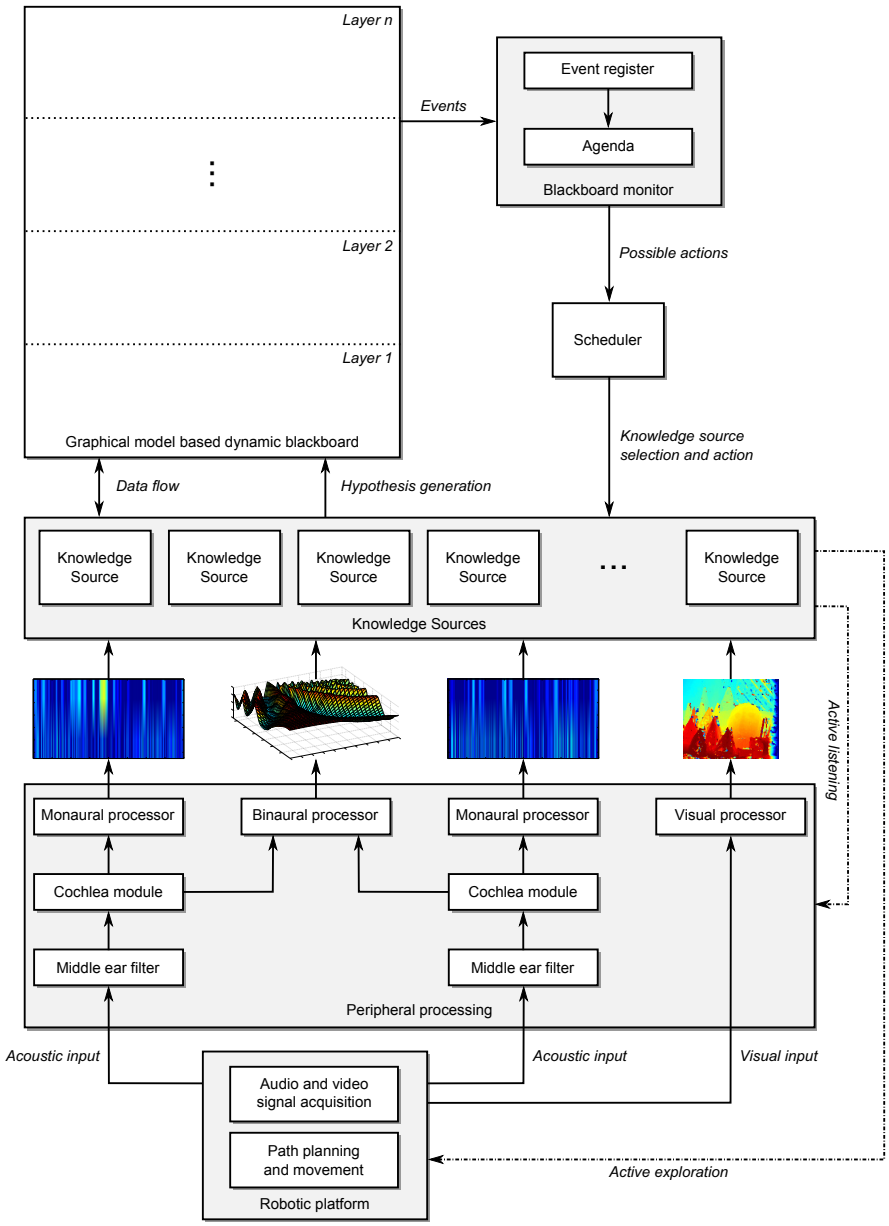

Figure 4.1: Overview of the general TWO!EARS software architecture.

A general overview of the TWO!EARS software architecture and the connections of the blackboard system to all other software modules is shown in Fig. 4.1. The blackboard system was released as part of the current TWO!EARS auditory model, in conjunction with the corresponding documentation[1] of all its software components.

1 `http://twoears.aipa.tu-berlin.de/doc/1.0/blackboard/`

4.2 Knowledge sources

In the search-and-rescue scenarios, there is a need for knowledge sources providing auditory and visual localization cues, for segmenting acoustic data, for general sound classification, visual object classification, for identifying the gender of speakers and speaker distress, and for recognition of keywords.

The quality-of-experience scenarios also require a localization knowledge source, and, in addition, they can profit from a musical genre recognition knowledge source. They will further rely on a knowledge source that is able to estimate the number of present sources and a knowledge source that can detect coloration.

In addition, the DASA-4 scenario also requires some blackboard-based movement control.

These necessary knowledge sources are specified in terms of their input variables, their functionality and their respective output below. For further information, please refer to the respective section in the official Two!Ears software documentation[2].

2 `http://twoears.aipa.tu-berlin.de/doc/1.0/blackboard/knowledge-sources/`

4.2.1 Localisation and Segmentation

Sound localisation from binaural cues

DNNLocalisationKS

- **Description**:
  Computes posterior probabilities of all azimuth locations from binaural cues using deep neural networks (DNN). The robot is assumed to be stationary.
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - AuditoryFrontEndKSs → 'KsFiredEvent'
- **Emits**:
  - 'KsFiredEvent' → SegmentationKS
  - 'KsFiredEvent' → ConfusionSolvingKS
- **Writes**:
  - 'locationHypothesis'
- **Reads**:
  - 'interauralCrossCorrelation'
  - 'interauralLevelDifferences'

Compute source localisation using sensorimotor flow

SensorimotorLocalisationKS

- **Description**:
  Computes the most likely azimuths (relative to the binaural head) of farfield sources on the basis of the analysis of the sensorimotor flow of the binaural head
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - AuditoryFrontEndKSs → 'KsFiredEvent'
- **Emits**:
  - 'KsFiredEvent' → SegmentationKS
- **Writes**:
  - 'locationHypothesis'
- **Reads**:
  - 'interauralTimeDifferences'
  - 'interauralLevelDifferences'

Source localisation by triangulation

EnvironmentalMapKS

- **Description**:
  Estimate positions of the acoustic sources by triangulation.
- **Interfaces**:
  - BlackboardSystem.robotConnect
- **Receives**:
  - MemoryFormationKS → 'TriangulationHypothesis'
- **Emits**:
  - no emissions
- **Writes**:
  - no information written
- **Reads**:
  - 'activateDisplay'
  - 'triangulatedLocations'

Computing visual categories and locations

VisualIdentityAndLocalizationKS

- **Description**:
  Computes the visual categories and locations of observed entities. Note that currently, the computation is based on pure simulation. This knowledge source may be limited to localization of humans.
- **Interfaces**:
  - BlackboardSystem.robotConnect
- **Receives**:
  - ReactToStimulusKS → 'KsFiredEvent'
- **Emits**:
  - 'KsFiredEvent' → AuditoryIdentityKS(speech)
- **Writes**:
  - 'visualIdentityHypotheses'
- **Reads**:
  - no information read

Forming audio-visual objects

BuildAudioVisualObjectKS

- **Description**:
  Forms basic audio-visual objects using blackboard information. Could also trigger feedback reactions if audio information contradicts visual information. This is a 'proof-of-concept' KS that has to be extended for more complex experiments.
- **Interfaces**:
  - BlackboardSystem.robotConnect
- **Receives**:
  - UpdateEnvironmentKS → 'KsFiredEvent'
- **Emits**:
  - no emission
- **Writes**:
  - 'audioVisualObjectHypothesis'
- **Reads**:
  - 'auditoryObjectHypothesis'
  - 'visualIdentityHypotheses'

Segmentation

SegmentationKS

- **Description**:
  Computes soft-masks based on binaural features that are modeled via circular distributions. Prior information about most likely source positions is gathered from location hypotheses that are generated by the LocationKS.
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - LocationKS → 'KsFiredEvent'
- **Emits**:
  - 'KsFiredEvent' → GenderRecognitionKS
  - 'KsFiredEvent' → KeywordRecognitionKS
- **Writes**:
  - 'segmentationHypothesis'
  - 'positionHypothesis'
- **Reads**:
  - 'locationHypothesis'
  - 'interauralCrossCorrelation'
  - 'interauralLevelDifferences'

Human visual localization

HumanVisualLocalizationKS

- **Description**:
  Localizes the position of a human standing upright perceived by a visual sensor. Returns a position on the ground plane, with respect to a defined world reference frame
- **Interfaces**:
  - BlackboardSystem.robotConnect
- **Receives**:
  - ReactToStimulusKS → 'KsFiredEvent'
- **Emits**:
  - 'KsFiredEvent' → VisualIdentityHypotheses
- **Writes**:
  - 'VisualIdentityHypotheses'
- **Reads**:
  - No information read

Object visual localization

ObjectVisualLocalizationKS

- **Description**:
  Localizes the position of a static object from a stereoscopic sensor. Prior information about most likely object positions (if detected also from audio data) is gathered from location hypotheses generated by audio based localization KS.
- **Interfaces**:
  - BlackboardSystem.robotConnect
- **Receives**:
  - ReactToStimulusKS → 'KsFiredEvent'
- **Emits**:
  - 'KsFiredEvent' → VisualIdentityHypotheses
- **Writes**:
  - 'VisualIdentityHypotheses'
- **Reads**:
  - No information read

4.2.2 Source classification

Sound classification

IdentityKS

- **Description**:
  Each instance of IdentityKS incorporates a model that generates hypotheses about the presence of a certain source-type in particular time span (extracted block from earsignals' streams). Many IdentityKSs can be instantiated – one for each type to be identified.
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - AuditoryFrontEndKS → 'KsFiredEvent'
- **Emits**:
  - 'KsFiredEvent' (upon generation of a new hypothesis)
- **Writes**:
  - 'identityHypotheses'
- **Reads**:
  - AFE features – depending on the actual model plugged in. Currently the models commonly use: 'ratemap', 'amsFeatures', 'spectralFeatures', 'onsetStrength'.

Gender classification

GenderClassificationKS

- **Description**:
  Generates hypotheses about the most likely gender of a speaker for a specific block of masked auditory features.
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - SegmentationKS → 'KsFiredEvent'
- **Emits**:
  - no emission
- **Writes**:
  - 'genderHypothesis'
- **Reads**:
  - 'segmentationHypothesis'
  - 'ratemap'

Speech emotion classification

EmotionClassificationKS

- **Description**:
  Generates hypotheses about the most likely emotion of a speech signal that is present in a scene.
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - Scheduler → 'AgendaEmpty'
- **Emits**:
  - no emission
- **Writes**:
  - 'SpeechEmotionHypothesis'
- **Reads**:
  - 'ratemap'
  - 'modulationSpectrum'
  - 'onsetStrength'
  - 'offsetStrength'

Keyword recognition

KeywordRecognitionKS

- **Description**:
  Computes the likelihood with which a segregated speech stimulus corresponds to each element of the set of trained keywords.
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - SegmentationKS → 'KsFiredEvent'
- **Emits**:
  - no emission
- **Writes**:
  - 'keywordIdentityHypothesis'
- **Reads**:
  - 'segmentationHypothesis'
  - 'ratemap'

Musical genre recognition

MusicalGenreRecognitionKS

- **Description**:
  Generates hypotheses about the most likely musical genre of a music signal that is present in a scene.
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - Scheduler → 'AgendaEmpty'
- **Emits**:
  - no emission
- **Writes**:
  - 'musicalGenreHypothesis'
- **Reads**:
  - 'ratemap'
  - 'modulationSpectrum'
  - 'onsetStrength'
  - 'offsetStrength'

Coloration

ColorationKS

- **Description**:
  Computes differences in timbre between a test stimulus and a reference stimulus after the model of Moore and Tan (2004).
- **Interfaces**:
  - BlackboardSystem.dataConnect
- **Receives**:
  - AuditoryFrontEndKS → 'KsFiredEvent'
- **Emits**:
  - no emission
- **Writes**:
  - 'colorationReference'
  - 'colorationHypothesis'
- **Reads**:
  - 'colorationReference'
  - 'filterbank'

Number of sources

NumberOfSourcesKS

- **Description**:
  Estimates the number of present sound sources.
- **Interfaces**:
  - BlackboardSystem.dataConnect()
- **Receives**:
  - AuditoryFrontEndKS → 'KsFiredEvent'
- **Emits**:
  - no emission
- **Writes**:
  - 'numberOfSourcesHypothesis'
- **Reads**:
  - 'interauralCrossCorrelation'
  - 'interauralLevelDifferences'

4.2.3 Movement Control

Turning to a perceived stimulus

TurnToKS

- **Description**:
  Rotates the robot towards a perceived stimulus, when the 'AuditoryObject-Formed' event is received.
- **Interfaces**:
  - BlackboardSystem.robotConnect
- **Receives**:
  - ReactToStimulusKS → 'AuditoryObjectFormed'
- **Emits**:
  - no emission
- **Writes**:
  - no information written
- **Reads**:
  - 'auditoryObjectHypothesis'

Computing direction of motion for audiomotor localization

MostInformativeLocalMotionKS

- **Description**:
  Computes the direction of the velocity vector of a binaural head which would locally improve the quality of the audiomotor localization of a single source
- **Interfaces**:
  - BlackboardSystem.robotConnect
- **Receives**:
  - ReactToStimulusKS → 'AuditoryObjectFormed'
- **Emits**:
  - No emission
- **Writes**:
  - No information written
- **Reads**:
  - 'auditoryOjectHypothesis'

Moving the robot to a given location

MoveToKS

- **Description**:
  Translates the robot towards a given location.
- **Interfaces**:
  - BlackboardSystem.robotConnect
- **Receives**:
  - MemoryFormationKS → 'InitRobotTranslation'
- **Emits**:
  - no emission
- **Writes**:
  - no information written
- **Reads**:
  - 'positionRequest'

5 Conclusion

This report contains the software specification for analyzing dynamic auditory scenes. After describing the implementation choices that have been made for the search-and-rescue and quality-of-experience tasks, it also defines the necessary knowledge sources for the blackboard architecture.

While it is not possible at this point to completely specify in detail all algorithms that will need to run on the blackboard system, it was possible to narrow down and specify those knowledge sources that will be necessary for solving the specified scenarios (DASA 1-4 and QoE 1-2). In line with the directory of work, we have thus provided all necessary specifications for the architecture, guiding developments for the final project year.

The operation of the scheduler has also been described in principle, but the interconnection of all elements on the blackboard will be a main target of research work in the coming year, holding significant scientific interest and allowing further optimization of the overall system performance based on the shared information on the blackboard storage. In this way, it will be possible to assess the value of a shared representation of acoustic scenes, which allows higher-level knowledge sources to utilize statistical knowledge from early preprocessing, but also informs preprocessing about higher-level expectations and thus provides many venues for interconnected optimization. The scenario descriptions presented here will provide a focus for evaluation of the system, and will play a key role in driving the development of the system forward.

Acronyms

DASA dynamic auditory scene analysis

KS Knowledge Source

QoE Quality of Experience

RMS root mean square

Bibliography

Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. (**2013**), "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language* **27**(3), pp. 621–633. (Cited on page 6)

Ma, N., Brown, G. J., and May, T. (**2015**), "Robust localisation of multiple speakers exploiting deep neural networks and head movements," in *Proc. Interspeech'15*. (Cited on page 12)

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 *EEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. (Cited on page 12)

Moore, B. C. J. and Tan, C.-T. (**2004**), "Development and Validation of a Method for Predicting the Perceived Naturalness of Sounds Subjected to Spectral Distortion," *JAES* **52**(9), pp. 900–14. (Cited on pages 12 and 24)

Raake, A., Wierstorf, H., and Blauert, J. (**2014**), "A case for Two!Ears in audio quality assessment," in *Forum Acousticum*, Kraków, Poland. (Cited on page 9)