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

Deliverable D4.3

Final integration-&-evaluation report

Ariel Podlubne, Benjamin Cohen, Chungeun Ryan Kim, Hagen Wierstorf, Gabriel Bustamante, Jens Blauert, Johannes Mohr, Jonas Braasch, Ning Ma, Patrick Danès, Sylvain Argentieri, Thomas Forgue, Thomas Walther, Tobias May, Yanmeng Guo, Youssef Kashef *

January 23, 2017

* 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:	4
Document number:	D4.3
Document title:	Final integration-&-evaluation report
Version:	1

Delivery date:	30th November 2016
Dissemination level:	Public (PU)
Nature:	Report

Editors:	Jens Blauert, RUB & Thomas Walther, RUB
Authors:	Benjamin Cohen–L'hyver, Chungeun Ryan Kim, Hagen Wierstorf, Youssef Kashef, Jens Blauert, Johannes Mohr, Jonas Braasch, Ning Ma, Sylvain Argentieri, Thomas Walther, Gabriel Bustamante, Patrick Danès, Thomas Forgue Ariel Podlubne, Tobias May, Yanmeng Guo
Internal reviewers:	Klaus Obermayer assisted by Christopher Schymura (Sec. 4.2.6)

Contents

1 Executive summary

It is well know that the human auditory system employs manifold feedback from higher to lower processing levels when forming auditory objects, analyzing auditory scenes and, consequently, generating perceptual worlds. This process incorporates input from other sensory modalities, such as vision and tactility. In consequence, auditory worlds are rarely purely auditive but to a certain extent multimodal. In the context of Two!Ears it is of major interest how such auditory and multimodal feedback can be computationally modeled, whereby these model should serve two purposes.

- To allow for experiments that access the effects that specific feedback loops might have on the auditory and cognitive processes that result in formation of perceptual worlds
- To explore possibility for the application of such feedback loops in technology

From the physiological literature it can be inferred that feedback loops may, in principle, originate from all stages of the auditory system up to the cortex, but only a limited number have yet been identified. However, this number is still too high for a three-year project like Two!Ears to cover all of them. Thus, we started with an extensive literature study and then, based on the results of this study (see Deliverable D4.1), set up a list of feedback loops to be actually modeled within our Two!Ears, namely, the *Priority List* shown on the next page.As will be reported in Chap. 4 of the current document in detail, all items from this list have been dealt with, although not all with the same depth. Looking at the results from the Two!Ears–participants' point of view, the following items are of particular relevance, regarding scientific as well as application progress. We thus consider them as the highlight amongst our project results as regards auditory and multimodal feedback.

- Turn-to reflex
- Sensorimotor feedback
- Signal-adapted control of ear filters
- Head turning modified by temporal salience
- Head turning modified by spatial salience
  - Localization accuracy
  - The Precedence effect
- Cognitive dynamic scene analysis and action control in search-and-rescue scenarios

The Priority List of feedback loops to be considered and explored

- **(A)** Olivo-cochlear reflex (MOCR)
  - **(a1)** Unilateral, contralateral and central control
- **(B)** Insertion of supplementary signal-processing units, triggered by decisions based on information taken from the blackboard
  - **(b1)** Specific enhancement filters, such as for male voice, female voice, baby voice
  - **(b2)** Precedence-effect processor
  - **(b3)** HRIR deconvolution
  - **(b4)** Dereverberation algorithm
  - **(b5)** Binaural noise-reduction algorithm
  - **(b6)** Machine-learned source identification: feedback-based selection of features and classifiers
  - **(b7)** Sensorimotor-cue processing
- **(C)** Cognitive-level feedback, for example, on the basis of labeled environmental maps as built from information taken from the blackboard and from experts
  - **(c1)** Interpretation of scenes and assigning meaning to their elements
  - **(c2)** Formation of attention and attention-based control of feedback processes
  - **(c3)** Performing quality judgments from the listeners' point of view, based on internal references
  - **(c4)** Initiating robot maneuvers for scene exploration, for example, for object–distance determination, approaching sources, triangulation
  - **(c5)** Keyword spotting
  - **(c6)** Requesting visual assistance through visual object localization and identification

2 Introduction and overview

There is strong physiological evidence that the auditory system provides various feedback in the course of auditory signal processing. Actually, top-down connections between almost all stages of auditory processing have been identified [1]. Further, it is apparent that the behavior of human beings, when actively exploring their surroundings, is significantly coined by what they hear while doing so.

Thus, it obvious that any endeavour to model the auditory system comprehensively must consider feedback paths as relevant elements. In the preparation phase of the TWO!EARS project, the prospective consortium had already thoroughly considered this necessity and drew up Fig. 2.1 as a concept for further discussion. Yet, because of its generality, this concept was not suitable as a guideline for goal-oriented scientific work.

Consequently, it was planned to set up a *Priority List* of feedback loops to be considered and explored in the framework of the project. Following intense literature studies, this list was released by the consortium after the first project year – see page 2 of the current document.

The items on the *Priority List* were selected with regard to the following criteria.

- What is scientifically of interest[2] and, in the same vein, relevant in terms of technological application
- What can realistically be achieved in WP4 in view of the duration (3 years) and person power (in summary, 2 scientists) of the project?

In this introductory section, we shall concentrate on selected project achievements regarding WP 4, that we deem particularly relevant. Details of all WP 4 result are reported in Chap. 4. In fact, all items of the *Priority List* have been dealt with.

1 See the introduction to Deliverable D4.1 for references to the literature

2 In the planning phase of TWO!EARS, there was hardly any scientific activity with the aim of including feedback loops in engineering models of the auditory system. However, this has changed in the meantime, obviously also triggered by the dissemination efforts of our consortium. Technological areas where this necessity has been recognized are, for instance, robotics and technical audiology (hearing instruments)

Feedback activity can roughly be grouped into *reflexive* and *reflective*. Reflexive feedback reacts fast and cannot easily be modified. Physiologically it is probably sub-cortically localized; in technological terms it can be conceived as hardwired. Reflecive feedback includes cognition[3], that is, symbol processing besides pure signal processing. This kind of feedback can usually be modified, for instance, learned and re-learned. Physiologically it is assumed to require cortical activity; in technological terms it contains software that can be modified.

- As regards reflexive feedback, we have worked on the so-called **Turn-to-Reflex**. This is the reflex of turning the head into the direction of the sound source in the case that a sudden sound pops up. The reflex helps putting the sound source into the visual field and thus identifying the reason for the sound. Interestingly, blind people do usually not turn the head in a frontal position to the source, but rather slightly to the side, namely, in the direction of best hearing. Whether this is still pure reflexive is worthy of discussion. In any case, appropriate head turning requires fast and reliable auditory source localization, also in situations with concurrent sources. In TWO!EARS we have mastered this task – Sec. 3.1

- If, in the course of sound localization, the commands that cause head movement, are considered in addition to the auditory cues, localization becomes more precise. It seems that as **sensorimotor-cue processing** happens subcortically, that it is a reflexive process – Sec. 4.2.7.

- In an interesting experiment, it was shown that the system can be made to "sharpen its ears" by **signal-adapted control of its earfilters**. To this end, the system tries to recognize whether incoming sound signals belongs to a classes of signals that it has learned beforehand. If this is the case, adequate enhancement filter are applied to the signals, to the end of enhancing binaural cues and, consequently, sound localization. As the system uses prior knowledge regarding the signal classes to consider, this a reflective feedback process – Sec. 4.2.1

- Head turning is not necessarily a purely reflexive process, but can also be controlled reflectively, based on the understanding of the scene concerned or at least of relevant parts of it. This raises the question of salience and attention. As an example, an algorithm for exploiting **temporal saliency** in the context of head movements has been developed in TWO!EARS. The basic idea is the following. If a sound source pops up in an environment for the first time, it is considered relevant and worthy of turning the head to. However, if a same or a similar (i.e. congruent) one appears, as compared to the prior one, it is considered less interesting as it is known already – and head turning may not be mandatory. Yet, if a new sound source enters the field

3 Goal oriented usage of knowledge and understandig

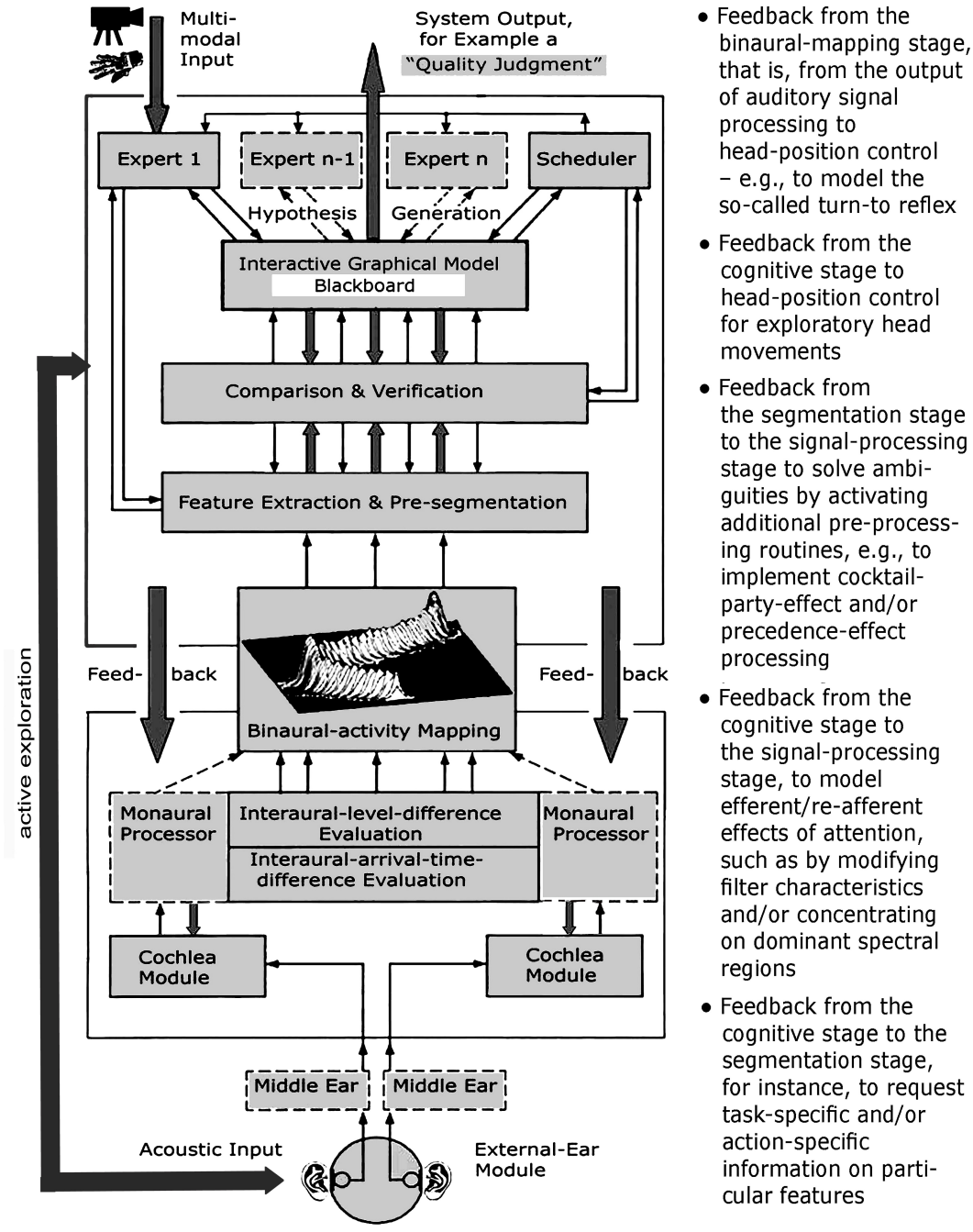


Figure 2.1: Conceptual plot of prospective feedback paths in binaural models [11]

that significantly differs from the prior one(s) (i.e. is incongruent), head turning is initiated again. The algorithm has been amended for including visual events in the object-building process (multimodal fusion) – Secs. 4.3.1 & 4.3.6

- As already mentioned above, *Localization accuracy* in single- and multiple source environments can significantly increased by reflective head movement. For example, it has been shown for two sources at different angles with respect to the listener that the classification of sound sources can be significantly improved. The best performing spatial arrangements have been found with a feature-based machine-learning approach

- A psychoacoustic phenomena that does not only comprise reflexive but also reflective elements, is the so-called **Precedence Effect** – Secs. 3.3 & 4.2.2. In general terms, it describes that in situations where a sound wave that comes directly from a source but is accompanied by delayed copies of it, such as wall reflections, the auditory event is formed in the direction of the first incoming wave front – the "direct sound". In Two!Ears, a novel, correlation-based Precedence-effect model has been developed, which can deal with ongoing sounds. The basic idea of the model can also be applied to HRTI-based signal enhancement, de-noising and de-reverberation – Secs 4.2.2–4.2.5. It could be demonstrated in Two!Ears, that latter two applications can be enhanced by informing the assessor visually about the size of the room in which he/she is sitting

- To simulate the behavior of robotic platforms with binaural sensors, a virtual environment with virtual objects has been developed, which can be set up and controlled by information rendered by the Two!Ears development system. For quick and reliable tests with cognitive functions of robots, it has however been realized, that the emulation of perceptual objects in the virtual environment is a versatile method to, for example, perform **cognitive dynamic scene analysis**. The emulation method has been verified with a virtual robot in a search-&-rescue (S-&-R) scenario (demo available). It is especially useful to investigate how humans and robots perform when actively exploring their environment and, while this process is under way, build an internal map of the scenario that they are in. The virtual robot is equipped with a camera and the virtual scene can be illuminated with varying intensity. In this way interaction of the auditory and visual modalities can also be studied – Chap. 5

The achievements of WP4 are thus focused on binaural perception from moving platforms. Most of them are scientifically original and have the potential of being be applied in the modern technology. The algorithms are well documented, and most of the software is open source. Details are communicated in the following Chapters.

3 Abstracts of demos for the final review

3.1 Sound localization with head movement

Tis demo concerns the use of head movements for sound localization is illustrated, with the robot in a fixed position. The robot does not restrict localization of sound sources to the frontal hemifield. Due to the similarity of binaural cues in the frontal and rear hemifields, front-back confusions often occur. Yet, human listeners rarely make front-back confusions because they use information obtained from head movements to resolve ambiguities in the binaural cues. To address this, the robot employs a hypothesis-driven feedback stage that triggers a head movement whenever the source location cannot be unambiguously estimated. One or more sound sources can be present about the robot. When a front-back confusion occurs, the robot actively rotates the head by a few degrees. Information before and after the head rotation is combined to help reduce front-back errors and to decide on the "true" positions of the sound sources.

3.2 Attention-driven sound localization

This demo demonstrates the effect of switching attention from one source to another during sound localization. Relatively few systems for machine hearing exploit top-down information in source localization, despite there being clear evidence for top-down (e.g., attentional) effects in biological spatial hearing. The robot has access to top-down models of the sources that are present in the environment. Information from source models is combined to improve the localization of the attended source by selectively weighting binaural cues. One target source and one interfering source are played from different locations around the robot. Top-down localization mechanism is activated or deactivated to show the beneficial effect of using source models. In particular, by switching attention between the two sound sources, the demo will show better localization of the attended source.

3.3 The Precedence-effect processor

The demo illustrates the capabilities of Two!Ears's Precedence-effect model in a video clip. For this purpose, we acoustically simulated the suite for the Two!Ears rescue scenario using ray-tracing. A video clip will highlight a virtual walk through the Two!Ears model suite. Based on the characteristic spatial/temporal pattern of direct sound from the source and early reflections, the model will determine the "correct" positions of different sound sources. Further, it determines with a top-down algorithms whether the sound source is occluded by a wall. The model resolves front/back directions via head movements, eliminate early reflections for further processing (e.g, speech recognition), and separates sound sources.

3.4 Robot "Jido" performing in an S-&-R scenario

To assess the capabilities of the *Bochum Experimental Feedback Testbed* (BEFT) in an emulated dynamic auditory scene, a simplified version of the ADREAM lab – which is the "computer apartment" set up at LAAS in Toulouse – has been replicated using the 3-D-modeling capabilities of *Blender* [12]. The emulation of the scenario has a duration of 400 s. It starts in normal lab conditions, but then, after $T_{event} = 60$ s, the situation evolves into a catastrophic scenario. After an assumed explosion, attendant lab employees turn into either victims or rescuers, and a fire starts in one corner of the lab. The robotic agent enters the scenario at $T_{start} = 0$ s and actively explores the terrain in order to infer the x/y-plane positions of all animate entities. With that done, the robot switches to idle mode. After T_{event}, the machine first triggers an audio-visual alert to warn potential co-workers, then it goes into rescue mode and evacuates all potential victims recognized in the environment. This procedure is illustrated in a video clip, where the screen capturing integrates a birds-view perspective of the scene, a 2-D environmental map, a log of all tasks addressed by the robotic device, and the images captured by a virtual camera attached to the head of the robot. For more details refer to Sec. 5.4.

3.5 Sensorimotor feedback

The reflexive motion of a binaural head which leads to the one-step-ahead most informative audio-motor localization of a source, that is, by means of *sensorimotor-cue processing*, has been implemented on the Two!Ears robot *Jido* – compare Sec. 4.2.7 for details of the process. A demonstration will be available at the final review meeting.

4 Items from the Priority List [1]

4.1 Olivo-cochlear reflex (A)

The medial olivo-cochlear (MOC) feedback has been implemented as a processor in the *Auditory Front-End* (AFE) of Two!Ears. This processor was intended to work in conjunction with the Dual-Resonance Non-Linear (DRNL) filterbank, also implemented within the AFE to correspond to the level-dependent nonlinear operation of the basilar membrane. The reflexive feedback was realized by using the output at the auditory nerve stage (ratemap processor in the AFE) to control the gain of the nonlinear path of the DRNL filterbank, through some internal adjustment to fit best the input-output relationship to the findings of [55]. The reflective feedback was realized by allowing for additional controls over the nonlinear path gain by external parameters as accessible from the blackboard system, for instance, by a Knowledge Source. More details of the operations and processing steps were reported in D4.2, Sec. 2.6.1. Details of the implementation in the AFE, including a running example, were reported in D2.3, Sect. 4.12.

4.1.1 Uni-lateral, contra-lateral and central control (a1)

Although this model was integrated and tested within the AFE, the lack of consistent evidence for functional benefits of using this feedback system, particularly for the conceived final scenarios, made the consortium decide to provide this feature as a potential testbed for future research. As described in D2.3 and D4.2, the modular nature of the software framework enables this DRNL-MOC processing chain to be activated interchangeably with the conventional Gammatone-filterbank-based processing chain. It is envisaged that the provision of this feature will extend the usefulness of the framework, as more findings become available through human tests against which the operation of this feedback mechanism can be validated further.

1 This list is project-internally known as the *Toulouse List*, because it was compiled after extensive discussion at a General Project Meeting in Sept. 2015 in F–Toulouse

However, this will certainly not be an easy task, as active olivo-cochlear feedback would add considerable nonlinearity and time-variance to the auditory front-end (AFE) and, if not properly controlled, may thus impair its capabilities to support the higher stages of the Two!Ears system. To investigate these issues in more detail makes an interesting topic for future work – for which the Two!Ears system provides an excellent basis and tool set.

Softwarestatus

Data/algorithm: partly (peripheral part of MOC reflex)
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: not intended

4.2 Insertion of supplementary signal-processing units, triggered by decisions based on information taken from the blackboard (B)

4.2.1 Specific enhancement filters, such as for male voice, female voice, baby voice (b1)

A number of psychophysical studies have found evidence for top-down effects in sound localisation. For example, covert shifts of attention can reduce reaction times when the spatial location of a target sound is cued by a preceding sound [81]. Physiological studies have also shown that sound localisation can be modulated by top-down influences. In the barn owl, sound localisation (including orienting behaviour such as head-and-body movements) is influenced by selective attention at the level of the midbrain, namely, responses associated with the position of behaviourally relevant stimuli (such as a food source) are enhanced [35]. Similarly, neural circuitry for gaze control exerts a top-down influence on the responsiveness of auditory neurons that are tuned to specific spatial positions [94]. Taken together, these findings suggest the existence of cross-modal mechanisms for top-down gain control of spatial hearing.

In contrast, relatively few systems for machine hearing exploit top-down information in source localisation. Two systems for binaural localisation of multiple sources recently proposed by [96, 95] use statistical frameworks to jointly perform localisation and pitch-based segregation. However, they do not use information about source characteristics other than through statistical models of pitch dynamics and binaural cues. A notable exception is the attention-driven model of sound localisation proposed by [56], in which top-down connections from a cortical model are used to potentiate responses to attended positions. However, attentional control in their model is driven by a simple neural circuit that fixates on sounds arriving from the same spatial position, and their approach is currently unable to localise multiple sound sources.

We propose a framework for sound localisation in which information from source models is used to selectively enhance binaural cues of the attended source. The system therefore combines top-down and bottom-up information flow within a single computational framework. We show that by exploiting source models in this way, sound-localisation performance can be improved under conditions in which multiple sources and room reverberation are present.

Estimating enhancement filters In the time-frequency (T–F) domain, the enhancement filters allow localisation cues that derive from a frequency channel dominated by the target source to be emphasised; or conversely, cues that derive from an interfering source can be

penalised. We propose an attentional mechanism that implements this idea. First, a T–F mask with the probability of each spectral feature of the observed signal being dominated by the energy of the target source is determined from prior models of the target and interfering acoustic sources. Then, the mask is employed during sound localisation to selectively weight the binaural cues.

First, let us denote by $\boldsymbol{y}_t = [y_{t1}, \ldots, y_{tD}]$ the spectral features (i.e. ratemap coefficients in log scale) extracted from the observed binaural signals at frame t. Similarly, we denote by $\boldsymbol{x}_t$ and $\boldsymbol{n}_t$ the spectral features for the target and mixture of interfering sources, respectively. In the log-ratemap domain, the relationship between these three quantities can be accurately approximated as

$$y_f \approx \max(x_f, n_f), \tag{4.1}$$

which is known as the *log-max* approximation [89, 75]. Note that, to simplify the notation, we have omitted the dependence of the spectral features on the time index, t.

Because x_f and n_f are unknown *a priori*, here we resort to a probabilistic approach for estimating the values of the T-F mask used by the attentional mechanism. Denoting the mask by $\boldsymbol{\omega}$, each of its elements, $\omega_f \in [0, 1]$ ($f = 1, \ldots, D$), indicates whether y_f is dominated either by the energy of the target source x_f and, hence, $\omega_f \approx 1$, or by the combined energy of the interfering sources, n_f, in which case $\omega_f \approx 0$. From a probabilistic point of view, and under the restrictions imposed by the *log-max* model in (4.1), ω_f corresponds to the following *a posteriori* probability,

$$\omega_f \triangleq P(x_f = y_f, n_f \leq y_f | \boldsymbol{y}). \tag{4.2}$$

To estimate this probability, we will employ statistical models describing the spectral characteristics of the sound sources. Let λ_s represent the spectral characteristics of a sound source, s, in a set of source models, $s = 1, \ldots, \mathcal{S}$. The set of source models are employed to jointly explain the observed ratemap features. In particular, given the observed log-compressed-ratemap feature vector, $\boldsymbol{y}_t$, extracted at time frame t from the binaural signals, we want to determine whether each feature, y_{tf}, is dominated by the energy of the target source, x_{tf}, or corrupted by the combined energy of interfering sources, n_{tf}. Under the *log-max* approximation [89] of the interaction function between two acoustic sources, ω_{tf} can be defined as the probability of y_{tf} being dominated by x_{tf} as follows.

$$\omega_{tf} = P(x_{tf} = y_{tf}, n_{tf} \leq y_{tf} | \boldsymbol{y}_t, \lambda_x, \lambda_n), \tag{4.3}$$

where λ_x and λ_n are the models for the target and interfering sources, respectively. Here, the source models, λ_s, are represented as GMMs with diagonal covariance matrices. λ_n is built by combining all the source models except that of the target source, that

is,

$$p(\boldsymbol{y}_t|\lambda_n) = \sum_{s\neq x} P(s) \sum_{m=1}^{M_s} P(m|\lambda_s)\mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{\mu}_s^{(m)}, \boldsymbol{\Sigma}_s^{(m)}\right), \tag{4.4}$$

where the prior probabilities of the sound sources $P(s)$ are assumed to be equiprobable. Alternatively, the above model can be expressed as a large GMM by pooling the Gaussians from all the source models together and multiplying the mixture weights by the prior probabilities of the corresponding sources, so that the resulting mixture weights sum up to one.

Using the expressions for the λ_x and λ_n models in (4.3), the final expression [75, 40] for the localisation weights ω_{tf} is given by

$$\omega_{tf} = \sum_{m_x,m_n} \frac{\gamma_t^{(m_x,m_n)} p_x(y_{tf}|m_x) C_n(y_{tf}|m_n)}{p_x(y_{tf}|m_x)C_n(y_{tf}|m_n) + p_n(y_{tf}|m_n)C_x(y_{tf}|m_x)}, \tag{4.5}$$

where m_x and m_n are the indices for the mixture components in the target source and interfering-sources models, respectively, p_x and p_n denote the Gaussian of the target and competing GMM models, and C_x and C_n are the corresponding Gaussian cumulative-distribution functions. The posterior probability, $\gamma_t^{(m_x,m_n)} \equiv P(m_x, m_n|\boldsymbol{y}_t)$, is defined as,

$$\gamma_t^{(m_x,m_n)} = \frac{p(\boldsymbol{y}_t|m_x,m_n)P(m_x)P(m_n)}{\sum_{m'_x,m'_n} p(\boldsymbol{y}_t|m'_x,m'_n)P(m'_x)P(m'_n)}, \tag{4.6}$$

and

$$\begin{aligned} &p(\boldsymbol{y}_t|m_x,m_n) \\ &\quad= \prod_f p(y_{tf}|m_x,m_n) \\ &\quad= \prod_f p_x(y_{tf}|m_x)C_n(y_{tf}|m_n) + p_n(y_{tf}|m_n)C_x(y_{tf}|m_x). \end{aligned} \tag{4.7}$$

Binaural-feature extraction An auditory front-end was employed to analyse binaural ear signals, consisting of a bank of 32 overlapping Gammatone filters with centre frequencies uniformly spaced on the ERB scale between 80 Hz and 8 kHz [91]. Inner-hair-cell function was approximated by half-wave rectification. Afterwards, the cross-correlation between the right and left ears was computed independently for each frequency channel using overlapping frames of 20 ms with a shift of 10 ms.

Two primary binaural cues, ITD and ILD, were extracted as features for binaural localization. The ITDs were estimated as the lag corresponding to the maximum in the cross-

correlation function output. The ILDs corresponded to the energy ratio between the left and right ears within an analysis window, expressed in dB. The ITD/ILD features were estimated for each frequency channel independently, forming a 2–D localization-feature vector, $\boldsymbol{o}_{tf}$, for time frame t, and frequency channel f.

Source spectral characteristics were modelled using ratemap features [19]. A ratemap is a spectro-temporal representation of auditory nerve firing rate, extracted from the inner-hair-cell output of each frequency channel by leaky integration and down-sampling – see Fig. 4.1. For the binaural signals used here, the ratemap features were computed for each ear and then averaged across the two ears. They were finally log-compressed to form 2–D feature vectors, $\boldsymbol{x}_t$.

Experiments The source model parameters were estimated from the ratemap features for each source separately, using the EM algorithm as described in Sec. 4.2.1. The training set included features extracted for each of the 72 azimuths considered in this study. Only "clean" features were used during the training stage, that is, the training signals were spatialised using the anechoic HRIR. The number of Gaussian-mixture components for each source was heuristically selected based on its spectro-temporal complexity as listed in Table 4.1. Speech material for the target source was drawn from the GRID corpus [25].

target	telephone	alarm	drums	symphony	baby	female
32	2	2	4	8	8	16

Table 4.1: The number of Gaussian-mixture components used for each source model

Each GRID sentence is approximately 1.5 s long and of the form "*lay red at G9 now*" spoken by one of 34 native British-English talkers. Here, the talker "s2" was used as the target source

Six types of sounds were selected as the interfering sources, with various amounts of spectro-temporal complexity. The ratemap representations of these interfering sounds are shown in Fig. 4.1. Their details are summarised below

1. **Telephone ring** Taken from Cooke's corpus [24], rhythmic narrow-band signals around 1 kHz and 3 kHz
2. **Alarm** Car alarm sound, rhythmic moderate-narrow-band signals between 800 Hz and 3 kHz
3. **Drums** Taken from [6], fast rhythms with clear energy onsets, synchronised across frequency, mostly overlapping regions of high speech energy
4. **Symphony** Mostly the sound of string instruments taken from Mozart's Symphony No. 40, 1$^{\text{st}}$ movement
5. **Baby crying** Less rhythmic and with higher formants than the target speech source

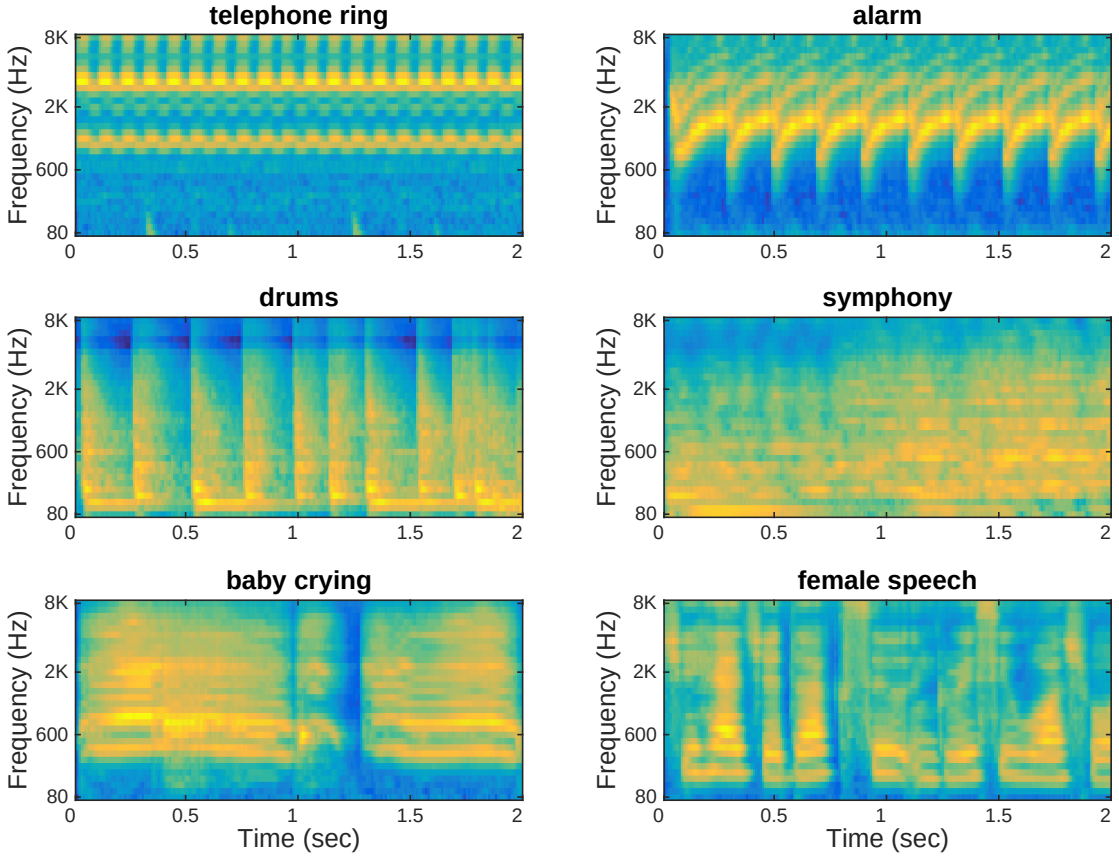


Figure 4.1: Ratemap representations of various interfering sounds used in this study

6. **Female speech** Taken from the TIMIT corpus [37], mostly overlapping the target speech frequency range

Fig. 4.2 (next page) shows examples of the estimated speech masks in the presence of the various noises. We also plot the oracle masks – speech/noise segregation using *a priori* information, to illustrate the quality of the estimated masks. Full evaluation of applying these masks to a localisation system is described in Deliverable D3.5, Sec. 4.1.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: no

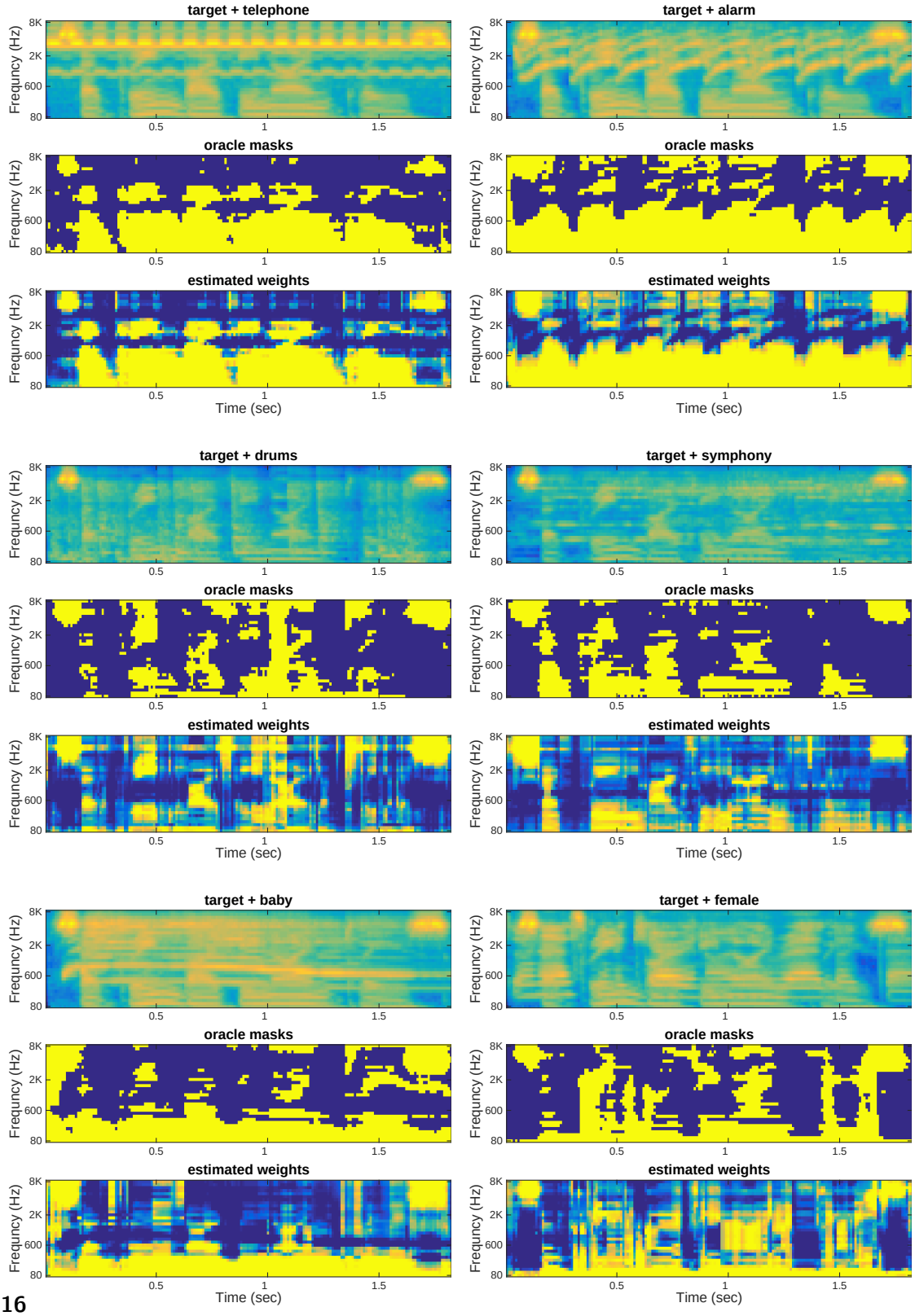


Figure 4.2: Estimated speech-segregation masks in the presence of various noises. **Yellow regions** are dominated by target speech. The oracle masks show speech/noise segregation using *a priori* information

4.2.2 Precedence-effect processor (b2)

Introduction The *Precedence Effect* describes the ability of the human auditory system to localize a signal in the presence of room reflections. The auditory system achieves this by disregarding or suppressing the localization cues of the reflected sound sources, hereby building on the fact that the direct sound always arrives first at the listener's ears, unless it is obstructed by an obstacle. For this reason, the Precedence Effect has historically also been named *Law of the first Wave Front.*

The ability to include processors to simulate the Precedence Effect is of fundamental importance for the success of the Two!Ears project, because the indoor robot rescue scenario we outlined in the original proposal relies on being also operationable in medium and highly reverberant environments. The Precedence Effect is also a great example of a top-down processing mechanism of the auditory system, which we sought to understand and simulate in Work Package WP3.

This section starts with an outline of the general Precedence-effect approach that we have taken. Then the steps are be discussed that were necessaryto expand the model from a system that can simulate laboratory studies with a single reflection to a system that can deal with physical environments, such as can be described by a complex impulse response.

The Two!Ears developments include

- The expansion of the model to deal with multiple reflections and diffuse reverberation
- The expansion of the model to generate binaural-activity maps, that is, plots depicting the position and arrival times of multiple early reflections
- The expansion of the model to handle head rotations
- The integration of the model into the Two!Ears framework

General description of the Two!Ears Precedence-effect model Figure 4.3 depicts the general structure of the Precedence-effect model. Parts of the model are similar to the models proposed by [10] and [15]. Basilar-membrane and hair-cell behavior are simulated with a Gammatone-filter bank and a simple half-wave rectifier. Thereafter, signals with very narrow bandwidths that do not evoke signs of localization dominance are sent directly to the cue-estimation stage as indicated by the switches in the model diagram. At such low bandwidths, a noise signal gives a periodic signal structure. Consequently, the decision of how to process a signal is made after analysis of the aperiodicity of the signal by using the YIN algorithm ([29]), labeled "YIN" in the model diagram. In the next stage, the missing half-waves are reconstructed (see boxes labeled "HR" in Fig. 4.3). The reason the signals are half-wave rectified in the first place is to demonstrate the proposed approach can operate using non-linear signals provided by the auditory periphery to the central

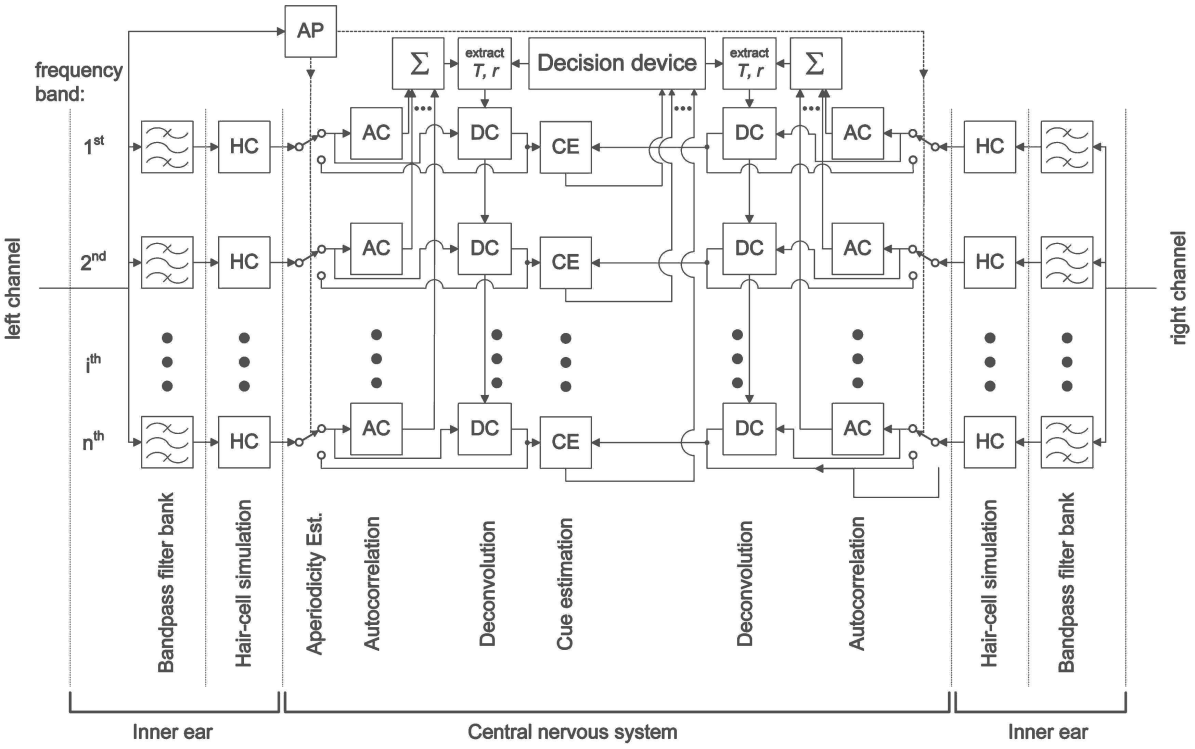


Figure 4.3: General model structure of the autocorrelation-based Precedence-effect model in its full implementation, including the simulation of the auditory periphery

nervous system, where the precedence mechanism resides. We have described the reliable performance of the bandpass filtering and half-wave compensation mechanism in [13] and thus do not include a detailed description in this report.

The autocorrelation functions are computed in each frequency band before they are integrated over frequency separately for the left and right channels (see "AC" labeled boxes in the diagram). The model then computes the lag delays, $T_{l,r}$, of the reflection measured from the arrival time of the direct sound source. Also, the Lag/Lead Amplitude Ratios (LLAR), $r_{l,r}$, are estimated from the integrated autocorrelation functions in the left and right channels. Both parameters are then used to adjust the lag-reduction filter (LR).

In the next stage, the interaural cues are estimated – boxes "CE" in Fig. 4.3. First, the normalized interaural cross-correlation functions are determined. After the lags are removed in the left and right channels based on the selected LLAR mode, ITDs and ILDs are computed in individual frequency bands. All localization cues are calculated using a running filter window – triangular shape, 50-ms duration. The window is moved forward

in steps of half the window length.

The estimated sound-source position is determined by combining the ILD and ITD cues. The decision device integrates both cues over time and frequency. Each value is weighted with the energy, $\sqrt{\hat{s}^2_{n,l} \cdot \hat{s}^2_{n,r}}$, in each time/frequency slot. Frequency weighting according to Stern et al. (1988) is also included in the model analysis. The stimulus lateralization, Ω, is given by the integrated cues according to [15].

The TWO!EARS Precedence-effect-model structure has two clear advantages over previous models, in particular, the ability to process ongoing sounds that humans preferably use for communication. In contrast, previous models have focused on the onset of impulses or other brief sounds. Another advantage of the model is that it actually removes the reflections. This way it can also be used to clean up signals for further processing, such as speech recognition.

In the next section, we will explain the fundamental mechanism of the Precedence-effect model before we get to the description of model expansions in the subsequent sections.

Fundamentals of the Precedence-effect model Although localization dominance is usually attributed to binaural effects, it is easier to first outline the model algorithm using a monophonic example of a direct signal, $s_d(t)$, and a single reflection, $s_r(t)$. Since the reflection is a delayed copy of the direct signal, we can write

$$s_r(t) = r \cdot s_d(t - T) \tag{4.8}$$

with the delay time T and the Lead/Lag Amplitude Ratio (LLAR) r. The latter can be treated as a frequency-independent, phaseshift-less reflection coefficient, given that the decrease in sound pressure with distance of lead and lag can be neglected. For a passive reflection, we typically find $r \leq 1$. At least this is the case when the direct sound source and the reflection are captured with a (hypothetical) omnidirectional receiver. In

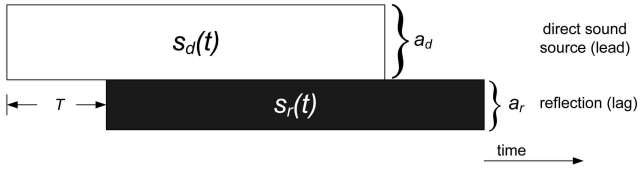


Figure 4.4: Time course of a single-channel lead/lag pair that consists of a direct sound source, $s_d(t)$, with amplitude a_d, and one reflection, $s_r(t)$, with amplitude a_r. The reflection is delayed by the delay time T

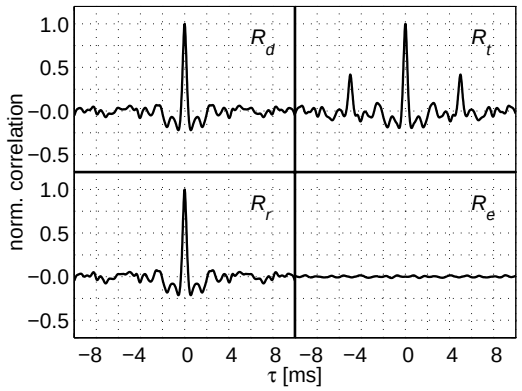


Figure 4.5: Autocorrelation functions for various broadband-signal configurations, that is, direct signal only, R_d, **(top-left panel)**; total signal, that is, direct sound with reflection, R_t **(top-right panel)**; reconstructed direct sound consisting of total signal with eliminated reflection, R_r **(bottom-left panel)**; error between reconstructed sound and its original, $R_e = R_r - R_d$ **(bottom-right panel)**. If applicable, the LLAR was set to 0.5, while the delay between lead and lag was set to 5 ms

the psychoacoustic literature, the delay time, T, is often referred to as the inter-stimulus interval (ISI). We can mathematically describe the total signal $s_t(t)$, which consists of the direct
sound, $s_d(t)$, and its reflection, $s_r(t)$, as follows – see also Fig. 4.4,

$$s_t(t) = s_d(t) + s_r(t) = s_d(t) + r \cdot s_d(t - T). \tag{4.9}$$

In the next step, we take the autocorrelation function of the total signal, from which we hope to extract information about both the delay time, T, and the LLAR, r, as follows,

$$\begin{aligned} R_{s_t} &= \int_{-\infty}^{\infty} s_t(t) \cdot s_t(t - \tau)\, dt \\ &= R_{s_d} + R_{s_r} + R_{s_d s_r} + R_{s_r s_d}\,. \end{aligned} \tag{4.10}$$

Aside from the two cross-correlation terms, we have two autocorrelation terms, one for the direct sound, R_{s_d}, and one for the reflection, R_{s_r}. In case that the direct sound is aperiodic, both functions should only have one peak, located at $\tau = 0$.

The top-left panel of Fig. 4.5 shows the autocorrelation peak for a broadband noise signal (direct sound only). The lead/lag condition is shown in the top-right panel of Fig. 4.5. Since, direct sound and its reflection are highly correlated to each other, we also receive two cross-correlation terms, $R_{s_d s_r}$ and $R_{s_r s_d}$. The first one has its maximum at $\tau = -T$,

the second one at $\tau = T$. Hence, we find for aperiodic signals,

$$R_{s_t} = \begin{cases} r s_d^2 & : \quad \tau = -T \\ \left(1 + r^2\right) s_d^2 & : \quad \tau = 0 \\ r s_d^2 & : \quad \tau = +T \end{cases} \tag{4.11}$$

The delay time between direct sound and reflection can easily be estimated by determining the position of one of the two side peaks. The next task is to determine the LLAR, r, from the ratio γ between one of the autocorrelation side peaks and the main autocorrelation peak, namely,

$$\gamma = \frac{R_{s_r s_d}}{R_{s_d} + R_{s_r}} = \frac{r s_d^2}{(1 + r^2)\, s_d^2} = \frac{r}{(1 + r^2)}\,. \tag{4.12}$$

By completing the square, we can resolve 4.12 for r as follows,

$$r = \pm\sqrt{\frac{1}{4\gamma^2} - 1} + \frac{1}{2\gamma}\,. \tag{4.13}$$

The ambiguities will be dealt with later in Sec. 4.2.2.

Lag removal through deconvolution Now that we know the delay between lead and lag, T, and the LLAR, r, we can design a simple filter that eliminates the lag from the total signal. Interestingly, the solution coincides with the impulse response of a cylindrical pipe resonator. The deconvolution filter, h_d, converges fairly fast and only a few iterations, N, are needed, thus we write

$$h_d = \sum_{n=0}^{N} (-r)^n\, \delta(T \cdot n). \tag{4.14}$$

Of course, in the ideal case N goes toward ∞. The mode of operation of the filter is fairly intuitive. The main peak of the filter lets the complete signal pass, while the first

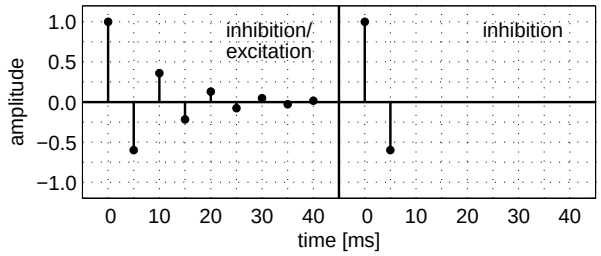


Figure 4.6: Impulse response of the new lag-suppression filter for eight iterations **(left panel)** as was applied to remove the lag in Fig. 4.5. The **right panel** shows the equivalent implementation for a simple approach with ipsilateral inhibition and no excitatory elements

negative peak is adjusted to eliminate the lag by subtracting a delayed copy of the signal. However, one has to keep in mind that also the reflection will be processed through the filter, and the second, negative delta peak will thus evoke a further signal component, which is delayed by $2T$ as compared to the direct signal. This newly generated component has to be compensated with a third, positive peak of the filter. A number of iterations are necessary to reduce the artifacts that result from the previous peaks. It is obvious that r cannot approximate one, since otherwise, the filter will not converge. For LLARs close to one, it is thus advantageous to limit r in (4.14).

In other current models, the delay between the first, positive and second, negative peak is typically set by a constant and not optimized for different stimuli – see Fig. 4.6. Also, the magnitude of the negative peak is set globally. In the case of our autocorrelation-based approach, the parameter of the filter is optimized for the amplitude ratio between lead and lag and the delay time between both. Further, the system response is no longer a plain inhibitory mechanism but rather one that includes both inhibitory and excitatory elements.

The bottom-left panel of Fig. 4.5 shows the autocorrelation result for a deconvolved signal. In this graph, the side peaks of the autocorrelation function, as are visible in the top-right panel, disappeared fully, and the function is very similar to the autocorrelation function of the lead only, plotted in the top-left panel of the same figure.

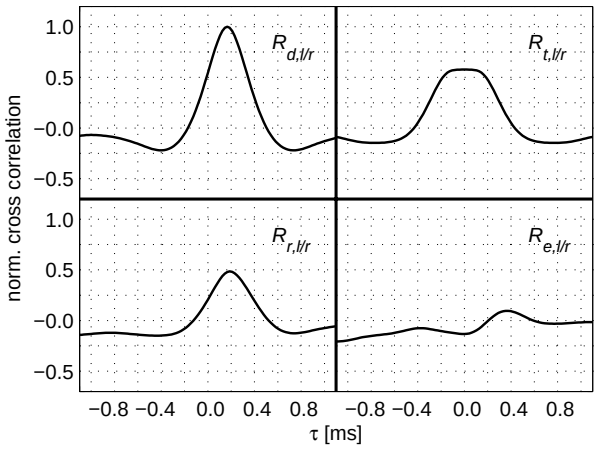


Figure 4.7: Interaural cross-correlation functions (ICC) for a binaural lead/lag pair based on ITD cues. The **top left panel** shows the ICC for the direct sound only, while the **top-right panel** depicts the situation for an additional lag sound. In the **bottom-left panel**, the lag has been removed from both channels using lag-suppression filters according to 4.14 before the ICC was calculated. The **bottom-right panel** shows the difference between the original ICC for the direct sound and the reconstructed one after lag removal

ITD-based signals Thus far, we have not specified a binaural mechanism to demonstrate localization dominance by any means, since we focused on a monaural algorithm to filter out the lag of the total signal. We continue to demonstrate this effect by applying our algorithm to a simple cross-correlation model for determining the interaural cross correlation (ICC) function, namely,

$$R_{s_l s_r} = \int_{-\infty}^{\infty} s_l(t) \cdot s_r(t-\tau)\, dt\,, \tag{4.15}$$

with the left and right ear signals, s_l and s_r.

A typical binaural lead/lag pair is created by applying an interaural time difference to the lead signal and processing the lag with a second ITD of the same magnitude but opposite sign. As the ISI is commonly defined as the delay time between lead and lag without considering spatial processing by applying ITDs, the actual delay time between lead and lag at both ear signals does not have to match the ISI completely. Usually, ITDs are applied such that the signal is preceded by half the ITD magnitude in one channel and delayed by the same value in the opposite channel. Accordingly, we estimate T and r for each channel individually, determine the deconvolution filters for both channels, deconvolve both channels separately, and then process the deconvolved binaural signal with the localization model. We can apply separate filters to the left and right ear signal before calculating the ICC.

Figure 4.7 depicts the calculated cross-correlation functions. The top-left panel shows the ICC for a single sound with an ITD of -330 μs (100-ms broadband-noise burst with a frequency range of 200–1000 Hz.) The position of the cross-correlation peak clearly indicates the ITD of the stimulus. In the top-right panel, the same sound is accompanied by a reflection (r=1, −330-μs ITD, 5-ms ISI). Then, we still observe a single peak being located in between the positions of both sounds. The bottom-left panel depicts the stimulus after the lag has been removed with the lag-suppression filter in both ear signals. Now the position of the ICC peak corresponds again to the ITD of the lag and, thus, the algorithm is demonstrating localization dominance. The bottom-right panel of Fig. 4.7 presents the negligible error between the original ICC function for the lead and the reconstructed function after lag removal, $R_e = R_d - R_r$.

To simulate data for cases with $r > 1$, as observed according to the Haas–effect, we need a method that can eliminate reflections for LLAR modes $r_{l/r} > 1$. To this end, we basically can use the lag-suppression filter of 4.14 for LLAR $r > 1$ as well. Since we only compute a limited number of iterations, we do not encounter the divergent properties of the filter. However, the filter needs to be scaled in amplitude and time, that is,

$$h'_d(t) = \frac{1}{r^{(N+1)}} h_d\left(t - (N+1)\,T\right)\,. \tag{4.16}$$

The first version of the Precedence-effect model needed a special mechanism to deal with reflection amplitudes that supersede the amplitude of the direct sound – described in detail in [13], see section on LLAR modes. With the expansion of the model, this mechanism is no longer needed and therefore not described in this report.

One form of displaying a binaural room-impulse response in line with our current model of the auditory processes in the central nervous system is the so-called *Binaural-activity Map*. The latter is a method to represent both the temporal and lateral positions of the reflected sound in a three-dimensional plot. One method of obtaining a binaural-activity map is to segment a BRIR into temporal slices of 10–100 milliseconds and then perform a binaural analysis over each time slice. Figure 4.8 shows an advanced method that can calculate a binaural-activity map from two running ear signals without the need of a measured impulse response. The algorithm is able to predict the delays and lateral positions of early reflections from the running ear signals. The algorithm cannot resolve the temporal pattern of diffuse reverberation. It appears to be the same case as with human listeners. This was revealed in a psychoacoustic experiment that was recently carried out within the TWO!EARS frame work [84].

Binaural-activity maps from Two!Ears test scenarios In this section, we simulate *Room-impulse Responses* (RIRs) for the rescue-scenario space of TWO!EARS. For this purpose, a ray-tracing method as was implemented. The Room-Impulse Responses were then convolved with voice signals (simulating the to-be-rescued inhabitants) and analyzed with the Precedence-effect model. The task of the model is to localize the inhabitants based on the position of the direct-signal source, and also to estimate whether the signal arrives directly at the robot, or if it is obstructed by an obstacle on its way.

In the course of the TWO!EARS project, we also investigated how the BiCAM localization model can be applied to a real-world scenario, for example, sound localization in an office suite. For this purpose, a ray-tracing model was implemented to generate binaural impulse responses for the binaural-model analysis. A geometrical model was defined as shown in Fig. 4.11, based on sound-reflecting walls, a source (red dot), and a receiver (blue dot). A set of rays is sent out from the sound source in every direction of the horizontal plane at equiangular distances of 5°. Each ray is then traced, and every time a ray meets a wall, it is reflected back using Snell's law, considering that the outgoing angle equals the incoming angle. The ray is traced until the 20^{th} reflection occurs, unless the ray exits the geometrical model. At every reflection, the sound level is attenuated by 2 dB across frequency to simulate acoustic wall absorption. The sound intensity is also attenuated over distance based on the inverse-square law, assuming the sound source to be of omnidirectional character. The collection of rays is shown in Fig. 4.10 as gray lines such that the rays become lighter in color with distance and decreasing sound pressure.

All rays are then collected at the receiver position assuming a 0.6 m-wide spatial window.

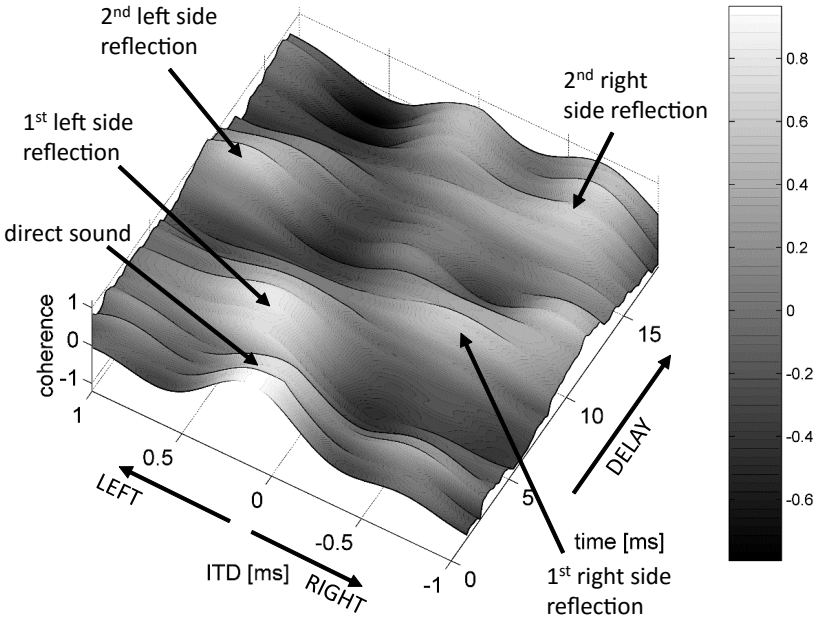


Figure 4.8: *Binaural-activity Map* extracted from a 22-s speech clip (spatialized with head-related transfer functions) located in the front (0-ms ITD, 0° Azimuth) and four early reflections arriving with delays of 6 ms (left side, 45° Azimuth corresponding to an ITD of +0.5 ms) and 9 ms (right side, 315° Azimuth, corresponding to an ITD of 0.5 ms), 12 ms (left side, 60° Azimuth, corresponding to an ITD of +0.7 ms) and 15 ms (right side, 300° Azimuth, corresponding to an ITD of 0.7 ms).

Each calculated ray is tested as to whether it intersects the spatial window at the receiver position. For each intersecting ray, it is then calculated how far it traveled from the source position to the receiver position, at which azimuth angle it arrives at the receiver position, and how many times it had been reflected (reflection order). Based on these data, a binaural room-impulse response is calculated in which a left/right HRTF pair is inserted at the correct delay and direction-of-arrival relative to the head. Each HRTF pair is calibrated to the amplitude that the ray should have – based on distance traveled and number of wall reflections. In addition, a late reverberation tail is generated at a constant level (assuming a statistically evenly distributed diffuse reverberation field) using an exponentially decaying Gaussian noise burst, adjusted to a reverberation time of 0.7 s. At the position shown in the left graph of Fig. 4.10, the diffuse reverberation level was about −10 dB lower than the combined level of the direct sound and the early reflections.

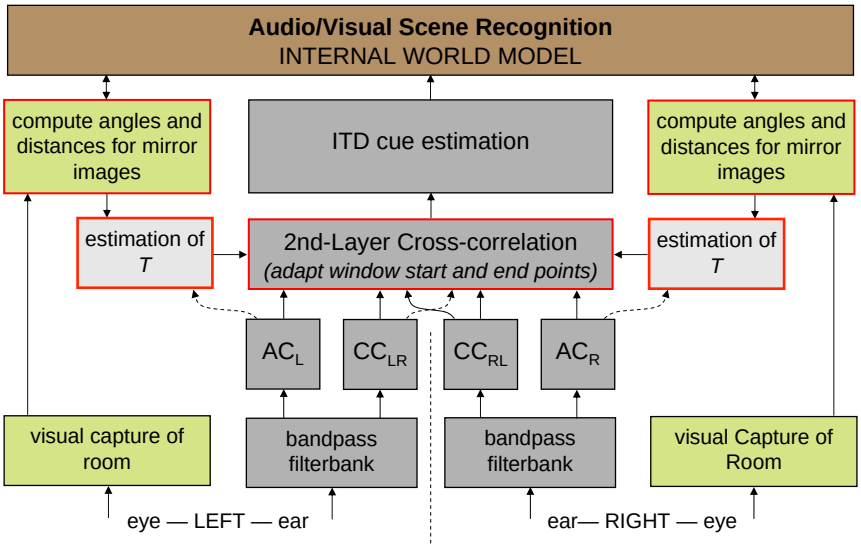


Figure 4.9: Architecture of a Precedence-effect localization model to demonstrate the build-up of the effect from visual cues using a combined auditory-visual model. The model calculates the delay, T, of the early side reflections from the visually-captured room geometry for the left- and right-ear channels. The estimated value determines the range over which the 2^{nd}-layer cross-correlation function will be performed

The results are then analyzed using the BiCAM (*Binaurally Integrated Cross-correlation/ Auto-correlation Mechanism)* algorithm [14] and a male speech sample from the Archimedes CD. The BiCAM algorithm was modified to transform the model's ITD estimates into azimuth angles using a remapping function according to [16] as shown in the binaural-activity map of Fig. 4.11 (top left graph). The plot shows the scenario in which the virtual head of the model is turned 30° away from the sound source, based on the scenario shown in the left graph of Fig. 4.10. Note that the data is presented in a room-coordinate system that faces the sound source directly. As can be easily seen, each time slice shows two ambiguous peaks, one for the front and one for the corresponding rear direction. This is the problem that was discussed in detail in [16]. In order to resolve the ambiguous peaks, the virtual head of the model is shifted by 60° to the opposite side – see the top-right graph of Fig. 4.11. The figure also displays the data in a room-coordinate system. Now we simply take the average of the two binaural-activity maps and, consequently, the ambiguous front/back-confusion peaks average out – see the bottom-left graph of the Fig. 4.11.

To demonstrate the effectiveness of the algorithm, the same scenario was simulated again, but this time with the virtual head facing the rear at 180° with temporal head-movement shifts to 150° and 210° to resolve front/back directions. It should be noted that there are

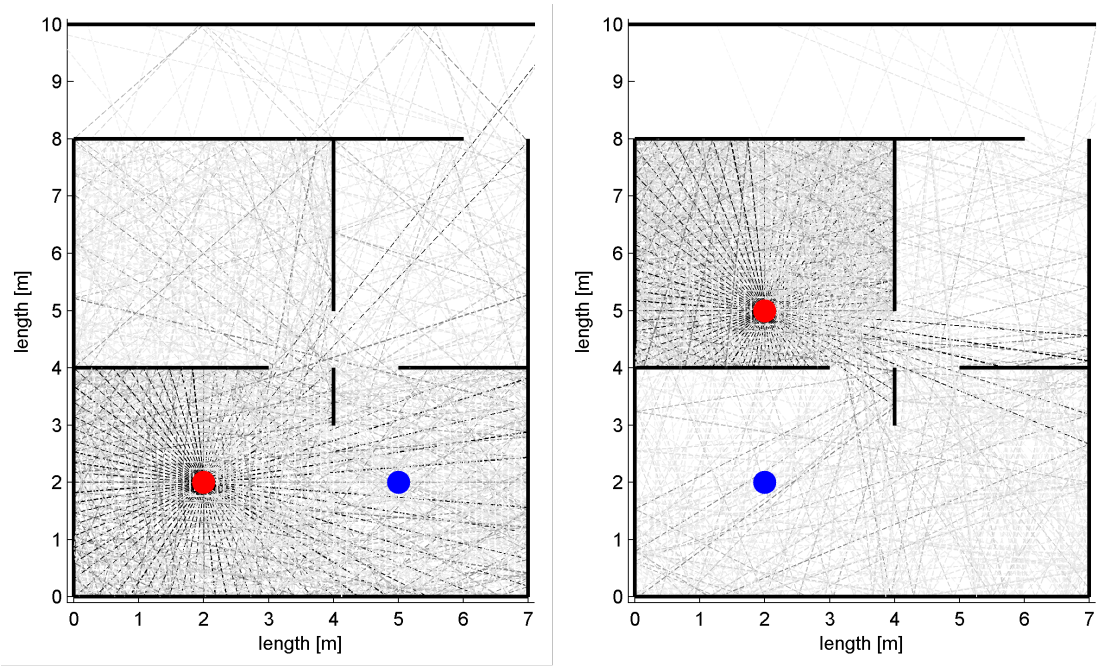


Figure 4.10: Ray-tracing simulation in a computer-generated office suite. **Left** Scenario 1 with a non-included sound source. **Right** Scenario 2 with an occluded sound source. Sound sources are depicted as a **red dot**, binaural receivers as a **blue dot**. The **gray level** of the rays lighten with decreasing distance and amplitude

two main differences between the model presented here and the model of [16]. Firstly, in the new model the head-movement algorithm is now applied to the estimated binaural-activity map and not to the binaural signal itself. This leads to the following two advantages. (i) The direct sound source angle can be computed separately from the early reflections, which results in a higher localization accuracy. (ii) The algorithm can also estimate the front/back direction of the reflections. However, the new model cannot yet calculate front/back directions from a continuously turning head like it is the case for the [16] model. The reason for this is that the time alignment method for the two autocorrelation functions currently requires a stable head orientation. Therefore, the new model calculates the front/back directions based on two distinct head positions until a better solution is found for the time-alignment method.

The analysis is concluded by computing a scenario in which the direct pathway between the source and the receiver is occluded by a wall as shown in the right graph of Fig. 4.10. Figure 4.12 shows the binaural-activity map for this case. While there is a distinct peak visible, the maximum correlation of 0.6 is much lower than was the case for the first scenario which yielded a maximum correlation of 0.9. Note that the binaural-activity

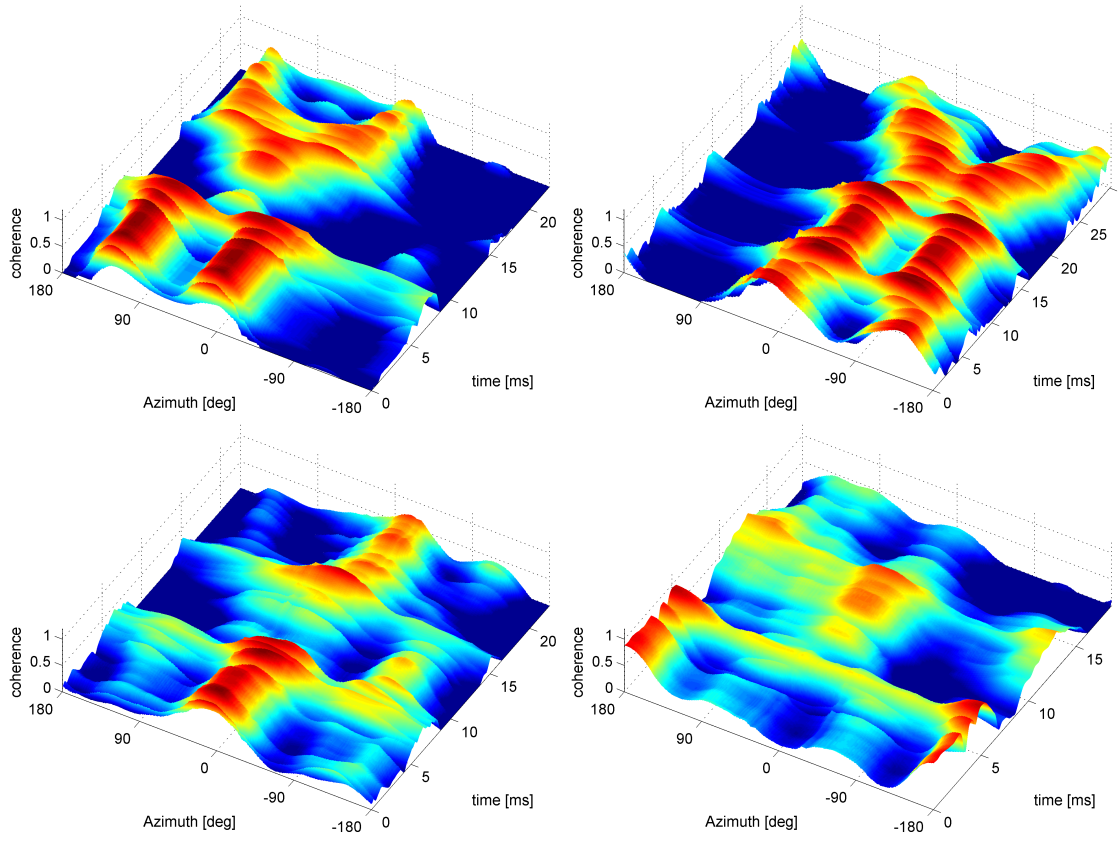


Figure 4.11: *Binaural-activity Maps* resulting from analysis with the BiCAM model [14], utilizing head movements. The **top-left graph** shows the results for Scenario 1 (Fig. 4.10) with the sound receiver pointing 30° left to the sound source (including 30°-head-movement compensation), the **top-right graph** shows the same condition but for the receiver pointing 30° to the right. The **bottom-left graph** shows the combined analysis to remove front/back confusions for a receiver pointing in the direction of the sound source (0°). The **bottom-right graph** shows the same condition, but for a receiver pointing away from the sound source (180°)

map was determined based on the average BiCAM analysis of 10 segments. Each segment by itself leads to a maximum coherence of one because the autocorrelation peaks have a main peak of one. However, in the occluded case, the outcome of the analysis is heavily influenced by the diffuse reverberant-signal component and the main peak averages out since its lateral position moves from segment to segment. In the case of Scenario 1, the binaural-activity map is stable from segment to segment and hardly influenced by the time-averaging method.

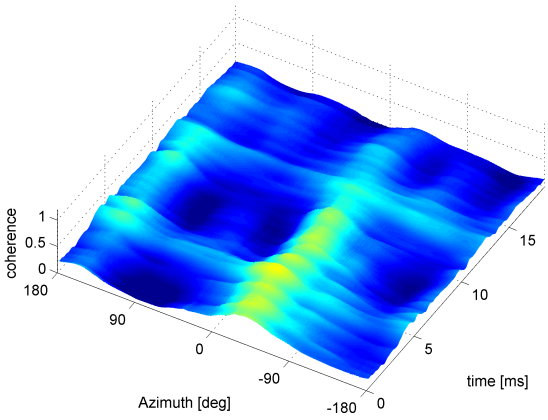


Figure 4.12: Binaural-activity Map resulting from BiCAM analysis, utilizing head movements for an occluded direct-sound source – simulating Scenario 2 depicted in the right graph of Fig. 4.10

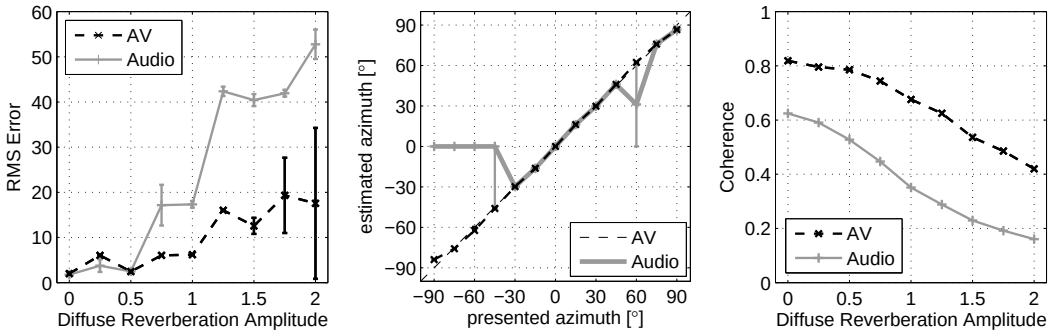


Figure 4.13: Localization results for the Precedence-effect model. **Left** Average root-mean-square (rms) error between the estimated and presented angles in degrees for the auditory (Audio) and the audio-visual model (AV) as a function of the amplitude of the late reverberation tail. The error bars show the standard deviation. **Center** Localization curves for the auditory and audiovisual models for a late reverberation amplitude of 1.2 s. **Right** Degree of coherence for the 2^{nd}-layer cross-correlation function for both models

Audio-visual Precedence-effect model The next goal was to use the visual model to improve the performance of our auditory Precedence-effect model. It is well known that visual cues about the position of sound sources can influence the auditory percept of the position of sound sources [48]. A visual model is used to provide information about the expected position of the early reflections to help the auditory model to disregard reflection-induced localization cues when estimating the direction of arrival of the direct sound from the source. Figure 4.9 depicts the architecture of the audiovisual model. As the figure shows, the visual model roughly determines the arrival times of the first two side reflections, measured with respect to the arrival time of the direct sound. For the estimation, the mirror-image technique [2] is employed. The information collected in this way is consequently used to optimize the performance of the auditory Precedence-effect model.

The auditory stages of the enhanced Precedence-effect model were adapted from BiCAM [14]. The model uses binaural ear signals to robustly localize a sound source in the presence of multiple reflections for the frontal horizontal plane. The core algorithm resembles a dual-layer spatio-temporal filter to separate auditory features for the direct and reverberant segments of a windowed signal to localize the direct sound source. At the first stage, the model separates the incoming binaural signal into auditory bands – see the two boxes in the center. Next, the model performs a set of auto-/cross-correlation analyses of the left and right ear signals within each auditory band – see the two boxes labeled "AC" and "CC".

A 2^{nd}-layer cross-correlation algorithm is then performed on top of the combined auto-correlation/cross-correlation algorithm. The underlying goal was to develop a method that incorporates the causality of the direct sound and its reflections. The conventional cross-correlation method does not reflect the temporal order of the incoming direct sound/ reflections, whereas the human auditory system takes this into account, as demonstrated by the Precedence Effect. The goal was achieved by introducing a second-layer cross-correlation analysis over the autocorrelation signal (e.g., R_{xx}) in one-channel and the cross-correlation signal (e.g., R_{xy}) in the second channel (see left graph, third row from bottom). The model compares the side peaks of both frequency-integrated functions (autocorrelation function and cross-correlation function). The side peaks for the left and right channels are correlated with each other and, by aligning them in time, the temporal offset between both main peaks can be used to determine the interaural time difference (ITD). In this way, the ITD of the direct sound can be found.

The visual input is used to determine a more suitable time window over which the 2^{nd}-layer cross-correlation algorithm integrates across the side peaks for the left and rights channels. In the standard version of the BiCAM model the 2^{nd}-layer cross-correlation is performed over the complete length of the positive side peaks (excluding the main center peak in the calculation) up to an internal delay of 120 ms, because it is unclear at which time delays

the early reflections arrive. In the audio-visual model we can use the roughly estimated delay times of the early reflections to reduce the duration of the integration window to focus only on early reflections and ignore the influence of the late reverberation tail. It should be emphasized that the visual model only provides input about the room geometry and not about the visual position of the sound source. A good example for such a scenario is a loudspeaker array in a room, where the listener does not know which loudspeaker produces the sound.

Stimuli The source signal used was a broadband Gaussian-noise burst with a duration of 6 s. The azimuth of the direct-sound source varied between the azimuth angles $-90°$ and $+90°$ in steps of 15°. The first early reflection was positioned at an azimuth angle of 0° (13-ms delay and an amplitude of 0.9 relative to the amplitude of the direct sound). The second early reflection was located at 45°, had a relative amplitude of 0.8, and a delay of 15 ms. In each case the direct-sound source and the reflections were spatialized using head-related transfer functions (HRTFs) from the MIT KEMAR database [36]. The late-reverberation tail consisted of an interaurally decorrelated, exponentially decaying Gaussian-noise burst with a reverberation time of 3.0 s and an initial time delay gap of 20 ms. The amplitude of the signal's late-reverberation component was varied between 0.0 and 2.0 in steps of 0.25, relative to the amplitude of the direct signal. A noise floor was added to the signal at -20 dB compared to the signal level. Each condition was tested 10 times with newly generated stimuli and noise floors.

Results The left graph of Fig. 4.13 shows the average root-mean-square (rms) error of the estimated azimuth versus the presented azimuth angles as a function of the amplitude of the late-reverberation tail. In all cases, the azimuth angles were computed from the estimated ITD cues for the direct-sound source using a remapping function as described in [16]. Without any late reverberation energy, the average rms errors are very small ($< 5°$). With a moderate increase of the reverberation-tail amplitude the error stays small until the diffuse-reverberation amplitude reaches a value of 0.75. Here the error for the auditory model increases to 17°, while the error for the audio-visual model still maintains the low error values. For both, the auditory-only and audio-visual model, the error increases nearly monotically with the diffuse reverberation amplitude, but the error of the audio-only model increases at a much higher rate, reaching an average error of over 50° at a diffuse-reverberation amplitude of 2. Comparatively, the error of the audio-visual model for the same condition is merely 18°. A big increase in error of the audio-only model can be seen between the 1.00 and 1.25 diffuse-reverberation-amplitude conditions. Although for most conditions the standard deviation of the averaged error is within a few degrees,. the standard deviation for the audio-visual model increases to 8° and 15° for the 1.75 and 2.00 diffuse-reverberation-amplitude conditions.

Next, we selected the 1.25 diffuse-reverberation-amplitude condition, that is, the position where the error increases substantially for the audio-only model, and plotted the median localization curves for this conditions – as shown in the center graph of Fig. 4.13. The audio-visual model accurately localizes the direct-sound source for all presented angles, and the errors are so small that the error bars showing the lower and upper quartiles are not noticeable. The audio-only model accurately localizes the signal for a few presented azimuth angles, that is, $-30°$ to $45°$, $75°$, and $90°$. Yet, for the remaining angles, the audio-only model is inaccurate by at least $15°$. Also, for the audio-only model, the errors are very small with exception of the $-30°$- and $60°$-angles, where the inter-quartile range is above $20°$. The right graph of Fig. 4.13 shows the average coherence values for the different diffuse-reverberation-amplitude conditions. Not surprisingly, the coherence values monotically decrease with increasing diffuse-reverberation amplitude for both the audio-only and the audio/video condition due to the decorrelation effect of the diffuse reverberation. It is important to note that the coherence values are consistently higher for the audio/visual model compared to the audio-only model, with an approximate difference of 0.2 between both cases.

Implementation The initial version of the novel Precedence-effect processor was implemented inside the Two!Ears's auditory-front-end framework (AFE), based on the work of [13]. In the presence of a binaural input signal with a delayed repetition or lag, the processor uses an autocorrelation mechanism and deconvolution to detect and remove the lag. Then it derives the ITD and ILD based on these lag-removed signals. The original processing algorithm as used in the work of [13] was revised and modified so that the processor could be implemented on the Two!Ears framework. The first implementation enabled chunk-based real-time processing, compared to the original version, which used the whole input-signal duration for the lag removal. The operation of this initial implementation as an AFE processor is described in D2.3, Chap. 4. A working demonstration of this stage was provided.

The processor subsequently underwent further modification, mainly to enable its use in conjunction with the DNN-based knowledge source (KS), which is part of the blackboard system. More specifically, the DNN-based localization KS requires cross-correlations (CCs) as well as ITDs and ILDs from the AFE for its operation. Since the first implementation of the Precedence-effect processor only supported ITDs and ILDs, the program code was modified such that cross-correlation information from the lag-removed input signals can be presented as an additional chunk-based output besides ITDs and ILDs. Therefore, a "Precedence-effect mode" of DNN-based localization was made available in which the *LocalizationKS* can request CCs, ITDs and ILDs from the AFE. This modification has been applied to the most recent Two!Ears software release, including the updated AFE-processor description.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: no

4.2.3 HRIR deconvolution (b3)

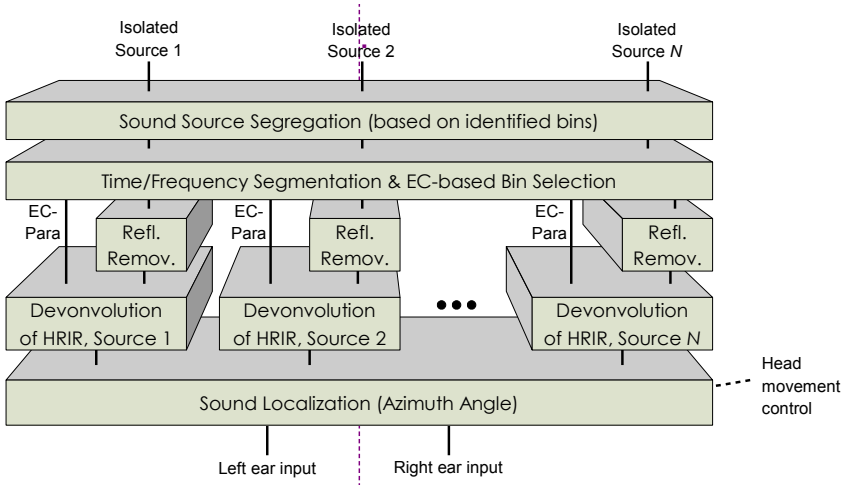


Figure 4.14: Architecture of the binaural model

This task was treated within a source-segregation framework that uses an *Equalization/ Cancellation* algorithm in the context of computational *Auditory-Scene Analysis* (CASA). The source-segregation algorithm builds on an HRIR-deconvolution mechanism to equalize the left and right signals in time and amplitude for each frequency band, such that they can be simply subtracted. The same source segregation algorithm will be used to demonstrate the outcome of task (b4), dereverberation algorithm – see next section.

The model, as depicted in Fig. 4.14, consists of three main stages, namely,

- A binaural Precedence-effect model to localize reverberant sound sources – see Sec. 4.2.2
- A mechanism to remove early reflections
- The source-segregation algorithm

Source-segregation model based on an equalization/cancellation process To create the cue-selection map, the left and right audio channels are sent through a Gammatone-

bandpass filterbank with 36 channels [67] and then segmented in time using a 512-point Hanning window with a step size of half the window length. Durlach's Equalization/ Cancellation (EC) model [32] was used to group the analyzed time/frequency segment to individual sources. This method was found to be more effective than the Faller & Merimaa method [33]. The EC method uses a null-antenna approach, considering that the lobe of the 2-channel sensor that the two ears represent is much more effective at rejecting a signal than filtering one out. In previous literature, the EC model is mainly used to explain the detection of masked signals. It assumes that the auditory system has mechanisms to cancel the influence of the masker by equalizing the left and right ear signals to the properties of the masker and then, after the equalization has been performed, subtract of the signal of one channel from that of the other one. Information about the target signal is obtained from what remains after the subtraction. For the equalization process, it is assumed that the masker is spatially specified by frequency-dependent interaural-time and -level differences as given by HRTFs. The two ear signals are then aligned in time and amplitude to compensate for these two interaural differences. The model can be extended to handle variations in time and frequency across different frequency bands. Internal noise in the form of time and amplitude jitter is used to degrade the equalization process to match human performance in detecting masked signals.

Figure 4.15 illustrates how this is achieved using the data in an auditory band with a center frequency of 750 Hz. For each graph, all possible ITD/ILD-equalization parameters were calculated using the method by [17], and the data for each bin shows the residual of the EC amplitude after the cancellation process. A magnitude close to zero (dark blue in the color map) means that the signal was successfully eliminated, because at this position the true signal values for the ITD (shown in the horizontal axis) and ILD were found (shown in the vertical axis). However, this is only possible for the left graph, which shows the case of an isolated target, and for the right graph, which shows the case of the isolated masker. In case of overlapping target and masker signals (as shown in the center panel) a successful cancellation process is no longer possible, because the EC model cannot simultaneously compensate for two signals with different ILD and ITD cues. As a consequence, the lowest point (aquamarine) with a coherence value of 0.15 is no longer sufficiently close to zero and, thus, the magnitude of the lowest coherence point can be used as an indicator of the presence of at least two overlapping signals in this time/frequency bin.

In order to construct a fully-capable *Cocktail-party processor* for CASA,[2] it is important to first analyze a cocktail-party scenario as an linear time-invariant (LTI) system. For a single talker, the signal sent to the ears can be represented as some speech signal, s_1, lateralized to some azimuth, θ, and convolved with the impulse response of a room, G.

2 A so-called "Cocktail-party scenario" is a scenario where multiple competing talkers are active simultaneously, and a listener tries to focus on one of them

This can be represented mathematically as

$$p_L(t) = s_1(t) * H_L(\theta_1, t) * G_L(t) \tag{4.17}$$
$$p_R(t) = s_1(t) * H_R(\theta_1, t) * G_R(t)\,, \tag{4.18}$$

with $p_{L,R}$ the signal measured at a given ear, s_1 the original speech signal desired for extraction, $H_{L,R}(\theta)$ the impulse response corresponding to a head-related transfer function (HRTF) for the signal at a given lateral angle, $G_{L,R}$ the binaural room-impulse response for the given ear. $*$ denotes mathematical convolution. Therefore, in order to extract the original target signal, the HRTF and room reflections must be removed.

For cocktail-party processing, this needs to occur for multiple speech signals. Such a system can be represented as

$$p_i(t) = \sum_{j=1}^{N} s_j(t) * H_i(\theta_j, t) * G_i(t)\,, \tag{4.19}$$

with p_i is the sound-pressure signal for a given ear, N the number of sources, s_j an individual speaker, θ_j the angle of the source, $H(\theta_j)$ the HRTF of the source at the lateralized angle, and G_i the binaural room-impulse response for a given ear. Intuitively, competing overlapping speakers cannot be trivially distinguished from one another. But the extraction is possible by using an adequate selection criterion [33] and by specifying the EC process only for the HRTF of the desired source to search for regions of energy containing *only* a target source.

While algorithms do exist to perform localization and anechoic signal extraction, no published binaural models exists that can perform sufficient reflection removal or source segregation for more than two sources. Multi-source binaural speech localization is presently still a challenge, with most published literature limiting models to extracting a single speaker in the presence of non-speech stimuli. Furthermore, as these systems are treated as linear and time invariant (LTI) and thus require a stationary head, traditional binaural models did not yet exploit head movements to increase their performance.

Source segregation In order to extract the target signal, the spatialized signals are segmented in both time and frequency to create individual time-frequency bins – similar to a short-time Fourier transform. The goal of the model is to compensate the target signal via division in the frequency domain. By dividing both the left- and right-channel short-time frequency spectra by the head-related transfer function for the angle of the target signal, the target is de-spatialized to the center, effectively creating a diotic signal in the bins for *only* the target. Thus, subtracting the right-channel spectrum from the left-channel one will leave zero residual energy, indicating that only the target is present in the given

time-frequency bin. However, if the masker is at all present in a time-frequency bin, then the two channels will not align perfectly and subtraction will yield some residual energy. The process can be modeled by the following set of equations.

$$x_1 = x_L * H_L^{-1}(\theta_{target}) \tag{4.20}$$

$$x_2 = x_R * H_R^{-1}(\theta_{target}) \tag{4.21}$$

$$E = \frac{\sqrt{(x_1 - x_2)^2}}{\sqrt{(x_L)^2 + (x_R)^2}}\,. \tag{4.22}$$

These energy values are used to generate a binary map. If the normalized energy contained in a bin is above a threshold (set to 0.85 for this model), then little cancellation occurred and the noise signal exists in that bin. The binary map is consequently set to 1 for these bins. On the other hand, if there was sufficient cancellation, then the target, i.e. the speech signal, is contained within the bin, and it is removed by setting the binary map to 0 for this bin. At this point, the binary map is then inverted such that $BM(t, f) = 1 - E$. This isolates the target source from the others, since the target is present in bins with little-to-no energy. and the masker is present in bins with lots of energy. Note that this allows for more than two sources to be run through the algorithm, and an individual voice can still be extracted. After applying the map to the signal via element-by-element multiplication, the signal is finally recombined in time and frequency, and only the target signal remains with the noise signal being removed.

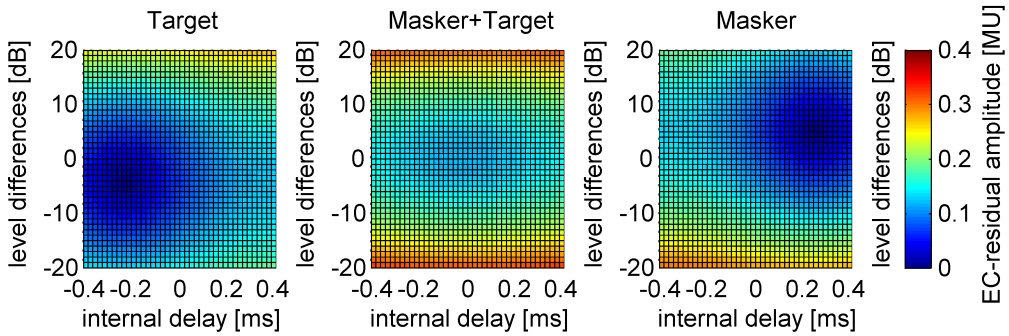


Figure 4.15: EC-model calculations for a broadband target/masker pair for one auditory band centered at 750 Hz. The **left graph** shows the target-only data. Note that the target location is at its minimum, close to zero **(blue zone)**, because the EC model was able to compensate for the target signal in this case, and no residuals remain other than internal-noise induced artifacts. The **center graph** shows the results for the combined target/masker presentation. In this case, the signal can no longer be fully canceled out, because the EC process can only eliminate one signal at a time. The lowest point (0.15 model units [MU]) is positioned in between the two locations of target and masker. The data for the masker are shown in the **right graph**

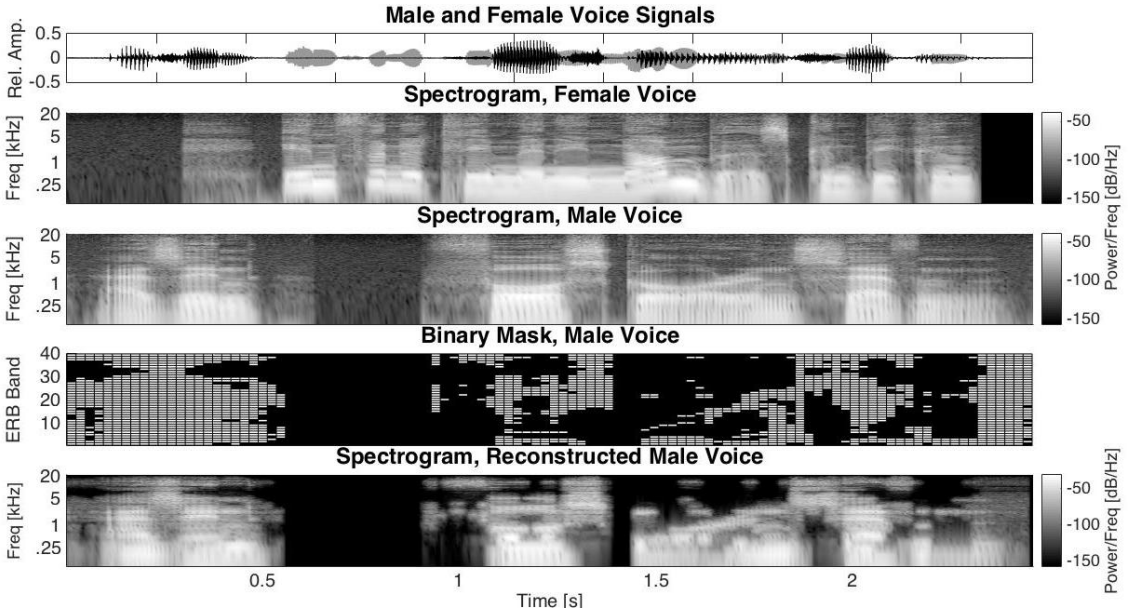


Figure 4.16: The EC process of the model. **1st panel** The male **(black)** and female **(grey)** voice signals. **2nd panel** & **3rd panel** Spectrograms of the female and male voice signals. **4th panel** The binary mask for extracting the male voice. **5th panel** A spectrogram of the reconstructed male voice

Results When segregation occurs with *a priori* knowledge of both source angles, the model successfully performs cancellation with high SNRs, depending on the angle between sources and head orientation. Background noise was injected to limit the performance of the model, as to be expected in realistic scenario. For this purpose, de-correlated pink noise was injected at approximately 20 dB-rms below the rms-pressure level of the masker.

The localization algorithm was found to be successful in localizing a single source in all presented cases. For multiple sources, the model must presently be aware of how many sources are present, however, the architecture of the allows the model set this by itself. This can be extended to estimate the number of sources. Furthermore, the localization algorithm can track moving sources with a high degree of accuracy. However, the case of resolving and extracting moving sources was not tested.

The output SNRs after head rotation showed maximum improvements ranging from 16 dB to 40 dB between different conditions across all angle pairs. This improvement was calculated by subtracting the SNR of the target-to-masker before head movement from the SNR of the same target-to-masker after head movement.

Discussion Without head movement, the model shows similar performance in SNR improvement as that of the binary-map model of Roman, Wang, and Brown [76]. Further, the model is able to successfully re-orient itself for maximimum cancellation with regard to the sources, and head movement results in significantly improved performance.

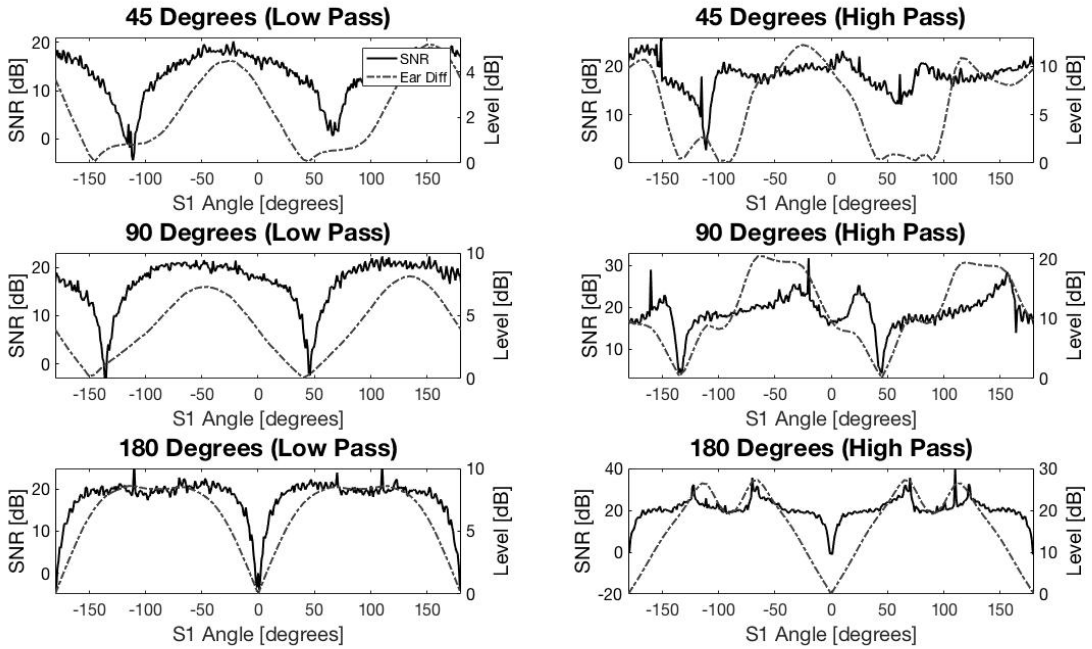


Figure 4.17: Performance of the model shown as signal-to-noise ratio or target/masker ratio after the target has been extracted from a target/masker mixture with an SNR = 0 dB – before signals are filtered with HRTFs. The results are shown as a function of head orientation. Each row depicts a specific source-angle condition between the masker and target. The **left graphs** show the results for signals low-pass filtered at 1.5 kHz, the **right graphs** show the results for signals that are high-pass filtered at 1.5 kHz, with the output further filtered to account for the ITD envelope. The **left-hand ordinate** depicts the SNR performance after cancellation, the **right-hand ordinate** shows the level of the target/masker pair for the right ear subtracted from the level of the target/masker for the left ear. Note that when both sources are directly in front of and behind the head, the model is unable to perform cancellation. These cases are recognized as undefined – marked 0 dB in the plots

Regarding the inversion of the binary map, this step is necessary to isolate the target speaker in the presence of multiple speakers. While the model was tested for only a single distracting signal, the inversion step theoretically allows the model to isolate the speaker in the presence of multiple noise signals. In doing so, the EC process is effectively using a *Null-antenna Approach*, considering that the lobe of the 2-channel sensor (representing the binaural hearing system) is more effective at rejecting a signal than it is at filtering and extracting a signal. For the case with only one noise source, this inversion step is not

required to segregate the target speaker. However, by applying the binary map specifically to regions where only the target exists, multiple noise sources can be canceled, leaving only the target in the residual energy.

To analyze the performance of the model after cancellation, the SNR – more precisely, the level of the target over the level of the masker (in dB) – as a function of head angle for a given fixed angle between sources for different azimuth angles between sources, is presented in Fig. 4.17. The solid black lines are modified filtered plots of the model's improvement in SNR (low pass filtered at 1.5 kHz on the left, high pass filtered on the right). This demonstrates the ability of the model to perform cancellation for the first source, placed at the angle corresponding to the independent axis, and the second source, placed at the sum of the angle of the first source and the angle in mentioned the title of each subplot. The shown trends demonstrate the model's ability to cancel is directly correlated with the signal-to-noise-ratio difference between the left and right ear – as shown by the dashed graphs. As such, the model demonstrates the better-ear or "squelch" hypothesis [18] in its ability to perform segregation. The head rotation of the model is capable of improving performance as it adjusts the system to operate at an angle where the better ear is at a distinct advantage.

So far, the model was only tested using anechoic stimuli. The presence of reflections is expected to interfere strongly with the localization and segregation of sources, because a significant quantity of phase information is impaired by reflections. Furthermore, while the model can segregate sound sources, recognizing or inferring the content of the speech signals was not attempted in Two!Ears.

Conclusions In summary, the proposed binaural model takes advantage of dynamic head movements to increase its performance. The model consists of three primary algorithms, namely, localization, head rotation, and segregation. By utilizing head rotation from the generated catalog and the inverse-filter method to create binary maps, the model can successfully isolate a target-speech signal from a mixture with a better performance than by simply using a binary map without head rotation. Model performance correlates with the better-ear hypothesis, and head rotation allows the model to take advantage of this method.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: not yet/can be made available
Runs on the robot: no

4.2.4 Dereverberation algorithm (b4)

Two different algorithms are used to de-reverberate signals in the context of sound-source segregation and preconditioning,

- Early reflections are removed using a reflection-removal filter
- Diffuse reverberation signal parts are removed along with other unwanted signals using time-frequency segmentation and grouping based on ideal masks

Currently, the reflection-removal filter has only been tested for one distinct reflection with constant ITD and ILD cues, but the method should work for HRTF-based reflections as well. The diffuse-reverberation removal works with HRTF-based signals.

Reflection-removal filter In our study, we have assumed that we are dealing with two stationary sound sources with simple broadband ITDs as localization cues. While the *Bi-CAM* localization model and the source-segregation model can handle sound signals that have been processed with head-related transfer functions (HRTFs), the current implementation of the reflection-removal filter cannot process HRTF-based stimuli yet. We have further assumed that the localization model can localize each of the two sound sources in isolation from one another and determine the values of the reflection amplitudes and delays for the left and right ear signals. In this study, each sound source has one early reflection, but the reflection parameters are different for both sound sources.

The early reflection is removed from the total signal prior to the application of the source-segregation algorithm – in Fig. 4.14. The filter design was taken from an earlier Precedence-effect model [13]. Instead of removing the early reflections independently for both sound sources, the reflection removal filter with the parameters for Source 1 was run over the total signal to isolate only Source 1, and the reflection-removal filter was run with the parameters for Source 2 across the total signal to isolate Source 2.

Test stimuli The examples shown in these sections were created using speech stimuli from anechoic recordings, taken from the “Music for Archimedes" CD. A a female and male voice at a sampling frequency of 44.1 kHz were mixed together such that the male voice was heard for the first half second, the female voice for the second half second, and both voices concurrent during the last 1.5 s. The female voice said: *Infinitely many numbers can be com(posed),* while the male voice said: *As in four, score and seven.*

For simplicity, the female voice was spatialized to the left with an ITD of 0.45 ms, and the male voice to the right with an ITD of -0.27 ms. In some examples, both sound sources (female and male voice) contain an early reflection. The reflection of the female voice is delayed by 1.8 ms with an ITD of -0.36 ms, and the reflection of the male voice is delayed

by 2.7 ms with an ITD of 0.54 ms. The amplitude of each reflection is attenuated to 80% of the amplitude of the direct sound.

For the examples that included a reverberation tail, the tail was computed from octave-filtered Gaussian-noise signals, windowed with exponential decay set for individual reverberation times in each octave band. Afterwards, the octave-filtered signals were added together for a broadband signal. Independent noise signals were used as a basis for the left and right channels and for the two voices. In our example, the reverberation time was 1 s – uniform across all frequencies with a direct to late-reverberation ratio of 0 dB.

Results for stimuli in the presence of diffuse reverberation In the next example, the EC model is used to determine areas in the joint time/frequency space that contain isolated target and masker components. In contrast to Fig. 4.15, the EC analysis is reduced to different ITD combinations and uses the second depicted dimension for time analysis instead of the ILD. Figure 4.18 shows the results for the EC-selection mechanism in the condition where the male voice is extracted. The top-left graph shows the selected cues (red areas). The color bar depicts the color code for the estimated values. These values correlate well with the male-voice signal shown in the sub-panel below (blue) but not with the female voice (red).

While the model also selects residual information from the female voice, most bins corresponding to the female voice are not selected (blue color). The top-right graph shows the binary mask that was computed from the left graph using a threshold of 0.75. The white tiles represent the selected time/frequency bins corresponding to the red areas in the left graph. The sub panel of the top-right graph shows the time series of the total reverberant signal (green curve, male and female voices plus reverberation), further, the isolated anechoic-voice signal (red curve), and the signal that was extracted from the mixture using the EC model (blue curve). In general, the model is able to both perform the task of isolation while noticeably removing the reverberation tail. The model still has issues with the onset of signals, presumably because the reverberation tail from the previous signal components interferes with the new signal components and de-correlates the overall signal. It also has problems processing segments with overlapping voices in time/frequency bins where the female voice dominates, or both signals are equally loud.

The bottom-left panel of Fig. 4.18 shows the same data as the top-left graph, but this time the EC algorithm targeted the parameters of the female voice. Now the algorithm primarily selects the time/frequency bins that corresponds to the female voice while correctly rejecting those that belong to the male voice. For both the male- and female-voice extraction examples, it is apparent that the algorithm currently performs much better for the mid-range and upper frequencies of the isolated signals parts. A fairly high percentage of time/frequency bins are missed when both signals overlap in the 1.0–1.5 s– time range,

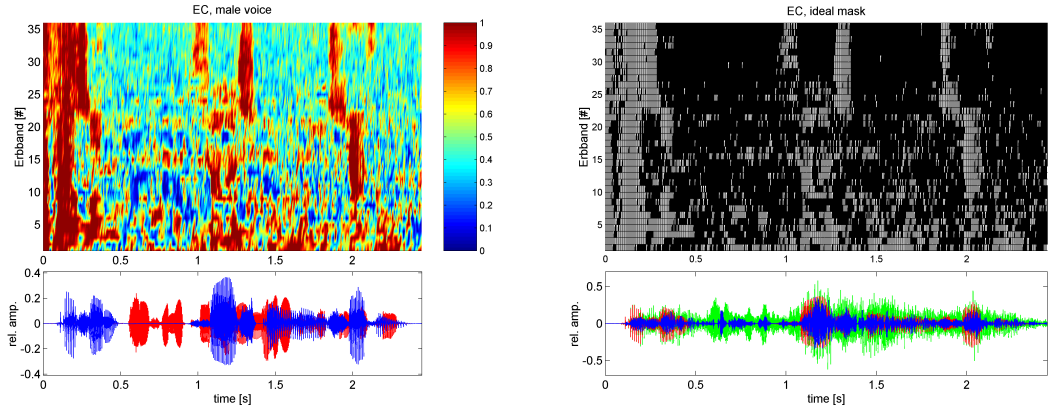


Figure 4.18: Signal-selection maps **(left graph)** and ideal masks **(right graph)** based on EC analysis to extract a male-voice signal in the presence of diffuse reverberation

while at low frequencies time/frequency bins are often over-selected.

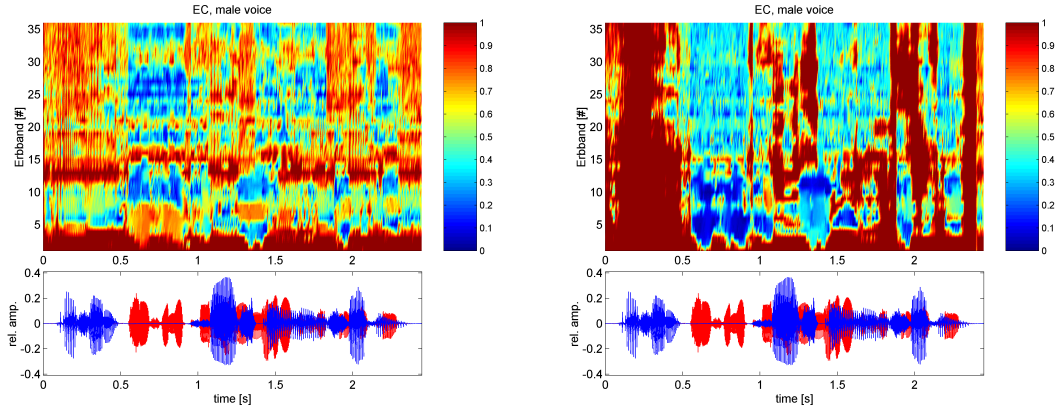


Figure 4.19: EC analysis **(left panels)** and binary maps **(right panels)** for the extraction of the male-voice signal from a female/male-voice sound mixture that contains early reflections, but no reverberant tail. The **left graph** shows the condition in which the early reflections were not removed prior to the EC analysis. The **right graph** shows the data where the early-reflection for the male-voice signal was removed The sub-panels show the male voice in **blue** and the female voice in **red**

Results for stimuli with early reflections Next, we studied how the algorithm can handle the removal of early reflections. For this purpose, we examined the test stimuli with early reflections as specified in the test-stimuli section (Sec. 4.2.4), but without a late

reverberation tail. Figure 4.19 shows the results of the procedure for the extraction of the male voice (top and center row) and female voice (bottom row). The method of data representation is identical to the one applied to Fig. 4.18, with the exception that in the previous figure a late reverberation tail was applied with no early reflections. The top-left panel shows the test condition where the early reflection of the male voice was not removed prior to the EC analysis. The analysis is very faulty. In particular, the signal is not correctly detected in several frequency bands, especially in the ERB bands #6 to #11 (220–540 Hz). At low frequencies, that is, bands #1 to#4, a signal is always detected and the female voice is no longer rejected. Consequently, the binary maps contain significant errors at the specified frequencies (top right graph), and the reconstructed male-voice signal does not correlate well with the original signal – compare the blue curve in the sub-panel of the top-right figure to the blue curve in the sub-panel of the top-left figure.

The two graphs in the center row of Fig. 4.19 show the condition in which the filter described in 4.14 was applied to the total signal in order to remove the early reflection for the male voice. Note that the female voice signal is also affected by the filter, but in this case the filter coefficients do not match the settings of the early reflection, because both the female and male voices have early reflection with different spatial properties, as would be observed in a natural condition. Consequently, the filter will alter the female-voice signal in some way but not systematically remove its early reflection. Since this signal is treated as background noise for now, its properties can be altered without worry as long as the signal characteristics of the male-voice signal can be improved. As the left graph of the center row indicates, the identification of the time/frequency bins containing the male-voice signal works better compared to the previous condition where no lag was removed – see Fig. 4.19 top-left panel. Note, in particular, the solid red block in the beginning where the male-voice signal is presented in isolation. This translates into a much more accurate binary map as shown in the right graph of the center row. There the extracted male voice signal (blue curve in sub-panel below) is more accurate than was the case in the last condition. Note that the amplitude of the extracted signal is lower than the male-signal component of the total signal – green curve. This is partially because the green curve contains reflections and thus greater overall energy, while the extracted signal does not.

It is important to emphasize that the application of the lag-removal filter with male-voice settings does not prevent correct rejection of the female-voice signal. Only in a very few instances does the model select a time-frequency bin with the female-voice-only region, that is, 0.5 s–1.0 s. The algorithm also does a much better job at extracting the male-voice signal from the mixture (1.0–2.5 s), than when no lag-removal filter was applied (compare top-right graph of the same figure). The bottom row of Fig. 4.19 shows how the lag-removal method works out for the female-voice-extraction process. Also for this case, the female voice can be extracted with the correct reflection-removal filter.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: installation in progress
Runs on the robot: no

4.2.5 Binaural noise-reduction algorithm (b5)

There are now plenty effective binaural noise-reduction algorithm available from literature, in particular, such as have been developed for advanced hearing aids . In Two!Ears we have thus decided to import a respective algorithm when needed, rather than develop an own one. A good introduction to the item can be found in [30], a more recent list of references is provided in [58].

Softwarestatus

Data/algorithm available: not yet
Code written and tested: not yet
Implemented on Two!Ears: not yet
Runs on the robot: not yet

4.2.6 Machine-learned source identification: feedback-based selection of features and classifiers (b6)

The contribution to this item was performed in the context of a study to investigate the role of of head rotation in sound-type classification. To this end, we used our *Auditory Machine Learning Training and Testing Pipeline* – see D 3.5, Sec. 3.3.1 – to investigate the following two questions, namely,

- How does the azimuth configuration influence sound-type-detection performance?
- Can a binaural robotic system improve performance by adequately turning its head?

This enables a setup which exploits the system's capability at characterizing the auditory scene in combination with its current state in order to form a top-down-feedback signal for modulating lower-level modules and tune them to optimize the system's performance in sound-type detection. Specifically, the system can make use of its sound-source-position estimates relative to its head orientation, and select a set of sound-type-classification models that are optimized for detecting sound types at this configuration.

The configuration of the scenes and features extracted to use in training and testing the models is described briefly. More details to the generation of auditory scenes and the

extraction of features are provided in D 3.5, Sec. 3.3.1. The optimum head orientation of different models that maximizes their performance is analyzed and the advantage of multi-conditional training over models that are specialized and tuned for specific orientation is discussed. We have found that for both multi-conditional and single-conditional models there is an optimum head orientation for each spread angle at which performance of sound-type detection is maximized.

Generation of auditory scenes We used the TWO!EARS binaural simulator to render four sets of binaural auditory scenes. The simulator convolves the audio source with an anechoic HRIR, measured with a Knowles Electronic Manikin for Acoustic Research (KEMAR), resulting in two-channel "ear signals" [92]. The following scenes are rendered.

1. Scenes composed of a single point source ("*target source*") emitting sounds from 12 different sound classes (including the general class) at 5 azimuth angles $\{0°, 22.5°, 45°, 67.5°, 90°\}$ – all mentions of azimuths are given with respect to the direction of the nose of the binaural head, see Fig. 4.20. These scenes are referred to as *clean sounds*

2. Scenes containing two point sound sources playing simultaneously a "*target source*", emitting sounds from all 12 classes, and a "*distractor source*" emitting only sounds from the general class
   a) Target- and distractor-sound sources were located at 17 combinations of azimuths: $\{0°/0°, 0°/45°, 45°/0°, 22.5°/-22.5°, 67.5°/112.5°, 0°/90°, 22.5°/112.5°, 45°/135°, 90°/180°, 22.5°/-67.5°, 45°/-45°, 90°/0°, 0°/180°, 22.5°/-157.5°, 45°/-135°, 67.5°/-112.5°, 90°/-90°\}$
   b) Target and distractor-sound sources were located at 16 combinations of azimuth angles: $\{0°/0°, 0°/45°, 45°/135°, 45°/-135°, -45°/-45°, -45°/-90°, 90°/180°, 90°/-90°, -90°/-45°, -90°/90°, 135°/-135°, 135°/45°, -135°/-90°, -135°/45°, 180°/180°, 180°/0°\}$

3. Scenes containing an ambient non-directional "*target source*" emitting sounds from all 12 classes
4. Scenes containing two ambient sound sources playing simultaneously: An ambient "*target source*" emitting sounds from all 12 classes and an ambient "*distractor source*" emitting sounds from the general class

Figure 4.20 displays our coordinate system in a top view, with the KEMAR head and two sources shown. The azimuth is always given with respect to the nose of the head (in clockwise direction), so in this example, we have an azimuth configuration of target and distractor source of 45°/135°. Sets 2.a and 2.b differ in the distribution of azimuths. In 2.a, the target source azimuth resolution was higher, but the target source was located between 0° and 90° in all 17 configurations – the distractor positions were distributed also

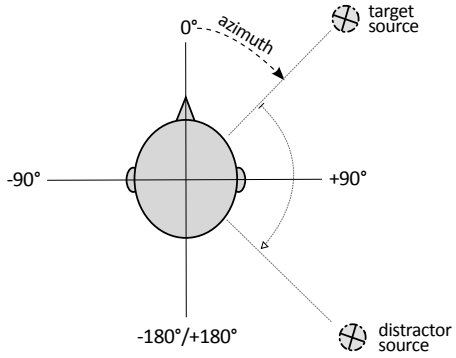


Figure 4.20: Top view of our coordinate system. Head and two sources are shown. Azimuth is always given with respect to the nose of the head (in clockwise direction). Target and distractor sources in this example are located at 45°/135°

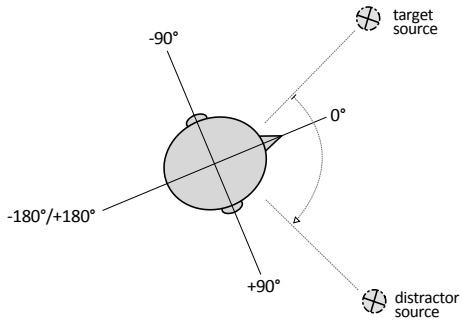


Figure 4.21: Compared to Fig. 4.20, the head has turned, and we now have a target-distractor azimuth configuration of −22.5°/67.5°

outside this quadrant. This set was used for all tests. In set 2.b, the target and distractor sources were distributed uniformly around the circle – at the cost of resolution. This set was used for training our multi-conditional models.

The "ear signals" resulting from the binaural simulation for each source were mixed at four different values of signal-to-noise ratio (SNR) {10 dB, 0 dB, −10 dB, −20 dB}. The "clean" sounds are assigned an SNR of ∞ dB. SNR was defined as the squared amplitude of the target source averaged over both the two binaural channels and time, divided by the averaged squared amplitude of the distractor source. The time-averaging only included times of sound activity and avoided periods of silence, which occur frequently in general environmental sounds.

Feature extraction The simulated binaural auditory signal was preprocessed by TWO!EARS's auditory front-end (AFE) to obtain representations using the following.

- **Ratemaps** Auditory spectrograms that resemble auditory nerve firing rates [39, 68, 26, 60]
- **Spectral features** 14 different statistics such as flatness, kurtosis, which summarize the spectral content of the *ratemaps* [69, 87, 49, 61, 54]
- **Onset strengths** Measured in dB for each time frame and frequency channel, calculated by the frame-based increase in energy of the *ratemaps* representation [52]
- **Amplitude-modulation spectrograms** Each frequency channel of the inner hair cell representation is analyzed by a bank of logarithmically-scaled modulation filters [62, 59]

All scenes were then decomposed into overlapping blocks of 500 ms. For each of our four target classes, we used the on- and offset times of the sound events to automatically label all blocks according to whether the target class was present (+1) or absent (-1) within the block. Blocks for which the target class occupied less than 375 ms were excluded from the analysis. We also dismissed blocks in which the distractor source energy was below −30 dB of the 99^{th}–percentile of the overall distractor-source energy. From the representations of each block we constructed two different types of feature vectors that were then used as input for the sound-type classification, namely,

- **Mean-channel features** Auditory front-end representations were averaged over the left and right channels. In addition, the first two deltas (discrete time derivatives) were calculated. Features were then constructed by computing L-statistics [3] (L-mean, L-scale, L-skewness, L-kurtosis) of representations and deltas over time
- **Two-channel features** Instead of averaging the signals over the two channels, the procedure for the mean-channel features was applied to each channel separately and then concatenated

Sound-type classification For each of our four sound categories (alarm, crying baby, female speech, and fire) we trained a binary classifier in a one-vs-all scheme, where each classification model decides whether a given auditory signal block contains a sound event from its respective target class. For classification, we used the *Least Absolute Shrinkage and Selection Operator* (Lasso) [85] – utilizing the "GLMNET" package [34, 74], a linear logistic regression model with an L_1 penalty for the regression coefficients. This penalty leads to sparser models by forcing many regression coefficients to zero. Therefore, Lasso is a classification method with an embedded feature-selection procedure. An important factor in determin-

3 L-statistics are given by L-moments, a sequence of statistics used to summarize the shape of a probability distribution [45]. L-statistics are shown to be more robust than conventional statistics, in particular with respect to the higher moments, and when a small amount of data is available – see Chap. 9 of [28]

ing the sparsity of the final model is the strength of the L_1 regularization term, which is controlled by the regularization parameter, λ. For adjusting its value, we performed a 7-fold stratified cross-validation on the training set for all 100 candidate values from the regularization path. We then chose the value with the best cross-validation performance and used it to train the model from the full training set. The full training set, using 75% of the sounds, amounted to roughly 75 k feature vectors and corresponding labels.

We considered two types of training schemes, namely,

- **Single-conditional (sc) training** The model was trained on data taken from one type of auditory scene and one condition only, i.e. one SNR and azimuth configuration – if not ambient
- **Multi-conditional (mc) training** The model was trained on data taken across auditory scenes and conditions

Evaluation Data were split into a training set for model building and a test set for estimating the generalization performance of the classifiers. In order to ensure that a block from the training set and a block from the test set never contained parts of the same sound file, training-test splits as well as cross-validation splits were conducted at the level of the original sound files. This means that the set of sound files for each class (including the general class) was randomly split into training set (75%) and test set (25%). Only the sounds from the training set were used to generate the auditory scenes for building the classification models, and only the sounds from the test set were used to generate the auditory scenes for evaluating the prediction performance. All models were built with three different training-test splits.

Performance was always evaluated on individual "single conditions", either

- On test data chosen from a scene and combination of SNR and azimuth configuration included in the training data –*iso-testing*
- On test data chosen from a scene and combination of SNR and azimuth configuration excluded from the training data –*cross-testing*

We used the *balanced accuracy* (BAC) as performance measure, which is defined as the arithmetic mean of sensitivity, i.e. the true positives (TP) divided by the size of the positive class (PC), and specificity, i.e. the true negatives (TN) divided by the size of the negative class (NC), that is,

$$BAC = \frac{1}{2}\left(\frac{TP}{PC} + \frac{TN}{NC}\right) . \tag{4.23}$$

Results We separately analyzed model performance for 16 of the 17 azimuth configurations in set 2.a of the binaural auditory scenes described in Sec. 4.2.6. These were assigned to three groups based on the "spread angle" between target and distractor source (α), namely 45°, 90°, or 180°.

Changing the azimuth configuration while keeping α fixed can be interpreted as a rotation of the head (see Fig. 4.21), if the distractor is always put in the same direction (clockwise or counter-clockwise) relative to the target. While this, at a first look, seems not to be the case for all our azimuth configurations, we could exploit the mirror symmetry between right and left hemisphere for our analysis. For instance, a model tested (and trained) at an azimuth configuration of ($t°/d°$) will perform equally to a model tested and trained at an azimuth configuration of ($-t°/-d°$).

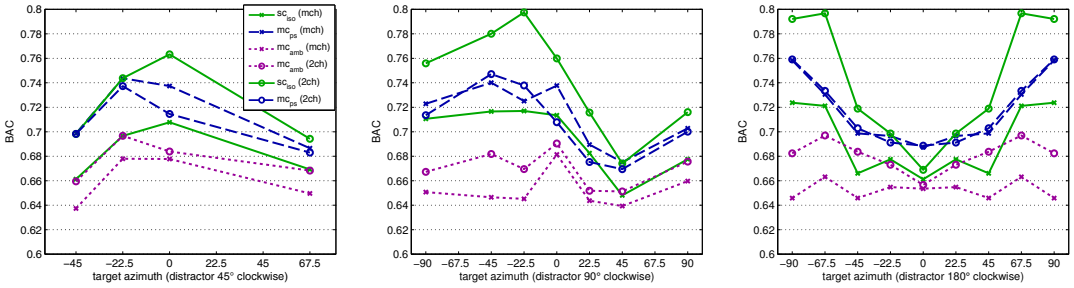
We trained the following different types of models.

- Multi-conditional models trained on ambient sounds (mc_{amb})
  - Models trained by combining data from the ambient source configurations from the sets of scenes 3 and 4 in Sec. 4.2.6 at all SNRs. These models cannot learn any directional information nor any head-related changes of the signals through the training data; and they cannot adapt to a particular SNR
- Multi-conditional models trained on point source sounds (mc_{ps})
  - Models trained by combining data from all point source configurations from the set of scenes 2.b in Sec. 4.2.6 at all SNRs. In this set, target and distractor source angles were uniformly distributed around the circle. These models cannot adapt to a particular SNR either, but get information about head-related changes of the signal through training data
- Single conditional iso-models (sc_{iso})
  - Models trained and tested under the same conditions, both with respect to SNR and azimuth

The single conditional iso-models models and the multi-conditional models were tested on the same single-conditional data from the set of simulated auditory scenes 2.a.

The results of these tests at an SNR of -20 dB are shown in Fig. 4.22, separately for spread angles $\alpha = 45°$ (a), $\alpha = 90°$ (b), and $\alpha = 180°$ (c). The x–axis denotes the angle of the target source. The distractor is always assumed to lie clockwise from the target at a relative angle of α. Color and line style indicate the different models, sc_{iso}, mc_{ps} and mc_{amb}, each on the mean-channel and the two-channel feature set indicated through different markers.

We find strong effects (up to around 13% difference between best-performing and worst-performing head orientation) with the following qualities.



(a) 45° spread **(b)** 90° spread **(c)** 180° spread

Figure 4.22: Performance of different models at -20 dB for different spread angles between target and distractor, averaged over data set splits and classes. Plotted over target azimuth with the distractor always put clockwise

- For the sc_{iso} and mc_{amb} models, the two-channel feature set models perform better than their mean-channel counterpart on all azimuth configurations, but the difference between the two varies greatly with head orientation and spread angle
- A higher performance can be reached with larger spread angle[4], although a saturation seems to be reached at $\alpha = 90°$. This follows the intuition that sources lying closely together are harder to discriminate
- The three spread angle groups show distinct performance profiles over the target azimuth
  1. For $\alpha = 45°$, best performance can be reached at a target azimuth of 0°, for $\alpha = 90°$ performance peaks at $-22.5°$ target azimuth (distractor at 67.5° – the situation depicted in Fig. 4.21), and for $\alpha = 180°$, we find the highest performance at a configuration with the target at $\pm 67.5°$. These are all configurations with the nose of the head being *close to the angle bisector* of target and distractor azimuths
  2. Sub-optimum performance is particularly found at configurations that put both target and distractor on one side of the head, which is well-observable in the 90° spread angle plot, and at configurations where one source is in the front and the other is in the back, which is well-observable in the 180° spread angle plot
- The performance differences between head orientations are stronger for the two-channel feature set (for sc_{iso} and mc_{amb} models). This is to be expected; it is more of a surprise that even the mean-channel feature set exhibits such a clear effect
- Although the mc_{amb} models did not learn any directional information during training

4 Results for the 0°/0°-condition, i.e. $\alpha = 0°$, are not shown in this figure, but performance is lower than for the other spread angles

and the mc_{ps} models had to learn from uniformly distributed azimuth configurations, the performance profiles of the three different model types for varying head orientations are, qualitatively, very similar

In summing up, this indicates that changes in head orientation have beneficial or detrimental effects on performance that are similar for all models, but that the effect size differs between models. Single conditional models that are specialized to the particular azimuth configurations can better exploit the spatial distribution of target and distractor at the beneficial head orientations than multi-conditional models. The point source multi-conditional models make use of this more effectively than the ambient multi-conditional models.

Conclusions Using our simulations, we investigated the effect of head orientation on sound type detection performance for different spread angles between target and distractor source. We found that for both multi-conditional and single-conditional models there is an optimum head orientation for each spread angle, at which the performance is maximized.

For the specialized single conditional models, the difference in performance between good and bad head orientations was more pronounced, but the performance profile over head orientations was qualitatively similar to the multi-conditional models. With the specialized models, the two-channel feature set showed a much stronger effect of head orientation than the mean channel one, since it can better utilize directional information.

In a binaural robot this might be used in an active listening system, if the system is able to reliably estimate the angular-source configuration of the environment. In particular, high-level control processes could turn the binaural head into an orientation that is likely to yield good performance. Therefore, the software for training and evaluating these models resides in the development system.

The evaluation was carried out on simulated data using the development system. Nonetheless, the software for deploying all trained source-type identification models is available in the TWO!EARS deployment system and available to run in simulation as well as on the robot. Software for switching between models specialized in specific head orientation is not available, that is, a knowledge source that makes use of the current head orientation and selects a model specialized in this orientation does not (yet) exist. Instead we opted to deploy a multi-conditional models whose performance is less sensitive to specific head orientations and thus mitigates the loss of performance due to wrongful selection of a model optimized for a particular head orientation.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: installation in progress

4.2.7 Sensorimotor-cue processing (b7)

Introduction Within the Two!Ears computational model of active auditory perception and experience, some feedback originates at the cognitive level and initiates context-dependent adjustment of bottom-up-processing functions and/or parameters. Further, hypothesis-driven activation of specific low-level processing procedures may be performed. Another kind of feedback operates at the sensorimotor level with no cognition between sensing and action and, thus, becomes effective on much shorter time scales. The *Turn-to-reflex* can be taken as an example in this regard.

Deliverable D4.2 proposes a mathematical statement to the synthesis of sensorimotor feedbacks which aim at improving the localization of a single source from a binaural head (Fig. 4.23), along the three-stage framework proposed in [21]. In brief, at each point in time, k, short-term directional cues are first extracted from the spectrograms, z_k, of the left and right signals over a small sliding window (Stage A) [71]. Then, up to time k, these cues are assimilated and combined with the motor commands of the head inside a Gaussian-mixture unscented Kalman filter (Stage B) [70]. This leads to the posterior-

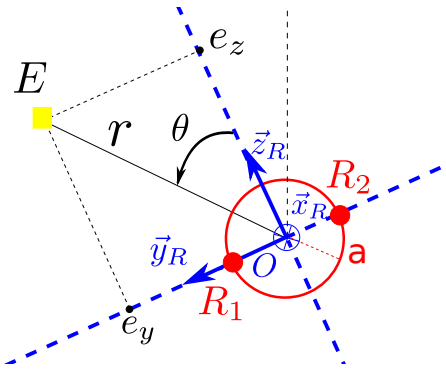


Figure 4.23: Sketch of the planar problem: The red circle depicts the binaural head, with O being its center and R_1, R_2 its left and right microphones, $\mathcal{F} = (O, \overrightarrow{x_R}, \overrightarrow{y_R}, \overrightarrow{z_R})$ its associated frame, and a the radius of an approximating sphere. The sound source, E, is located on the horizontal plane $(O, \overrightarrow{y_R}, \overrightarrow{z_R})$, at the distance r and azimuth θ. $\overrightarrow{z_R}$, $\overrightarrow{y_R} = \frac{\overrightarrow{R_2R_1}}{\|\overrightarrow{R_2R_1}\|}$ and $\overrightarrow{x_R}$, respectively, point frontwards, leftwards, and downwards

probability-density function (pdf) or "belief",

$$p(x_k|z_{1:k}) = \sum_{i=1}^{I_k} w_k^i \mathcal{N}(x_k; \hat{x}_{k|k}^i, P_{k|k}^i), \tag{4.24}$$

of the head-to-source relative position, $x_k = (e_y, e_z)^T$, at time k, where $z_{1:k}$ is a shortcut for $z_1, \ldots, z_k$, and $w_k^i, \hat{x}_{k|k}^i, P_{k|k}^i$ term the weight, mean and covariance of each hypothesis. Front and back can be disambiguated, and both the source range and azimuth can be recovered. Given the initial belief (4.24) extracted from sensorimotor flow up to time k, sensorimotor feedback (Stage C) aims at controlling the subsequent motion of the binaural head in the "most informative" way. In other words, that sequence of motor commands, $u_k, \ldots, u_{k+N-1}$, is sought for, which, on average, maximizes the spatial information with respect to the source as contained in the N-step ahead belief, $p(x_{k+N}|z_{1:k+N})$. Then, u_k is applied, leading to the new hidden state, x_{k+1}, and, consequently, a new sequence of Stages A, B, C follows.

The solution developed in D4.2 relies on the following assumptions.

- The information criterion to be maximized is the expectation over the next N unknown measurements, $z_{k+1:k+N}$, of the entropy, $h(x_{k+N}|z_{1:k+N})$, of the N-step ahead belief, $p(x_{k+N}|z_{1:k+N})$, that is,

$$J_N(u_k, \ldots, u_{k+N-1}) = \mathsf{E}_{z_{k+1:k+N}|z_{1:k}}\big[h(x_{k+N}|z_{1:k+N})\big] \tag{4.25}$$

$$= \mathsf{E}_{z_{k+1:k+N}|z_{1:k}}\Big[\mathsf{E}_{x_{k+N}|z_{1:k+N}}\big[-\log p(x_{k+N}|z_{1:k+N})\big]\Big] \tag{4.26}$$

- The exploration is guided by a scalar directional cue, z_k, which solely depends on the source relative azimuth, $\theta_k = -\text{atan2}(e_y, e_z)$. The observation model takes the closed form

$$z_k = l(x_k) + v_k = \bar{l}(\theta_k) + v_k, \;\; z_k \in \mathbb{R}, \;\; v_k \sim \mathcal{N}(0, R_k) \tag{4.27}$$

  with v_k the measurement noise and R_k its (co)variance. For instance, $\bar{l}(\theta_k)$ can express the Woodworth–Schlosberg ITD approximation between antipodal microphones over a spherical head for a farfield sound source [1]

- The initial belief, $p(x_k|z_{1:k})$, is reduced to a single Gaussian pdf

$$\hat{p}(x_k|z_{1:k}) = \mathcal{N}(x_k; \hat{x}_{k|k}, P_{k|k}) \tag{4.28}$$

  for instance, by keeping the most probable hypothesis of $p(x_k|z_{1:k})$, by turning $p(x_k|z_{1:k})$ into its Gaussian-moment-matched approximation, or by computing the moment-matched approximation of the most probable "branch" – that is, set of con-

tiguous hypotheses with similar azimuths of $p(x_k|z_{1:k})$

- The one-step-ahead solution is sought for, namely, $N = 1$. Then, under the above assumptions, the optimum rigid motion, u_k^*, to be applied to the sensor between times k and $k+1$, is shown to satisfy

$$u_k^* = \arg\min_{u_k} J_1(u_k), \text{ with } J_1(u_k) = \mathsf{E}_{z_{k+1}|z_{1:k}}\big[h(x_{k+1}|z_{1:k+1})\big], \tag{4.29}$$

$$= \arg\min_{u_k} h(x_{k+1}|z_{1:k+1}), \tag{4.30}$$

$$= \arg\max_{u_k} h(z_{k+1}|z_{1:k}), \tag{4.31}$$

with $h(x_{k+1}|z_{1:k+1})$ the entropy of the next filtered-state pdf, $p(x_{k+1}|z_{1:k+1})$, and $h(z_{k+1}|z_{1:k})$ the entropy of the predicted measurement pdf, $p(z_{k+1}|z_{1:k})$. $h(z_{k+1}|z_{1:k})$ comes as an increasing affine function of the logarithm of the variance, $S_{k+1|k}$, of $p(z_{k+1}|z_{1:k})$. Importantly, given the initial belief, $\hat{p}(x_k|z_{1:k})$, $F_k(u_k) = \log S_{k+1|k}$ can be computed inside an unscented Kalman filter. Though $F_k(u_k)$ may be involved, an approximation of its gradient, $\nabla F_k(u_k^0)$, at any u_k^0 can be derived by means of successive first-order Taylor expansion and the unscented transform ([50]), leading to

$$F_k(u_k^0 + \mathrm{d}u) = F_k(u_k^0) + \nabla F_k(u_k^0)^T \mathrm{d}u, \tag{4.32}$$

with $\mathrm{d}u = (\mathrm{d}T_y, \mathrm{d}T_z, \mathrm{d}\phi)^T$ the infinitesimal motion vector applied around u_k^0. Both the state- vector posterior-covariance matrix (that is, the covariance of $p(x_{k+1}|z_{1:k+1})$) and the predicted measurement variance, $S_{k+1|k}$ – obtained by the unscented filter transform – do not depend on z_{k+1}, hence the simplifications in (4.30– refeq-J1-c)

In view of these assumptions, the following heuristical rules of thumb are shown to be in effect for the definition of the one-step-ahead optimum motion, u_k^*, where $\mathcal{E}_k$ stands for the 99%-probability confidence ellipse associated with the initial belief $\hat{p}(x_k|z_{1:k})$:.

- The control input, u_k^*, must orient [resp. must not orient] the auditory fovea (that is, front direction) [resp. the interaural axis] of the binaural head towards $\mathcal{E}_k$
- The control input, u_k^*, must drive [resp. must not drive] the center of the binaural head on the (line supported by) the minor axis [resp. the major axis] of $\mathcal{E}_k$
- The control input, u_k^*, must drive the binaural head closer to $\mathcal{E}_k$

As shown in Fig. 4.24, these guidelines make $\mathcal{E}_k$ intersect as many "iso-z loci" (that is, loci of the head-to-source positions, x, corresponding to given values of z in the absence of noise) as possible, what in turn increases the entropy, $h(z_{k+1}|z_{1:k})$. They lead to a shrinking of the confidence ellipse associated to the next filtered-state pdf, $p(x_{k+1}|z_{1:k+1})$, after the incorporation of $z_{1:k+1}$ or, equivalently, to the decrease of $h(x_{k+1}|z_{1:k+1})$.

Information-based one-step-ahead-based optimum motion Considering the notations introduced in Fig. 4.23, let $u_k = (T_y, T_z, \phi)$ be the sequence of finite translations, T_y, T_z, along axes $\overrightarrow{y_R}, \overrightarrow{z_R}$, followed by the rotation around $\overrightarrow{x_R}$ to be applied to the binaural head between times k and $k+1$. Denote by

$$\mathcal{T} = \{T \triangleq (T_y, T_z) \in \mathbb{R} \times \mathbb{R} \,|\, {T_y}^2 + {T_z}^2 \leq r_{max}^2\} \text{ and } \mathcal{R} = \{\phi \in \mathbb{R} \,|\, |\phi| \leq \phi_{max}\} \quad (4.33)$$

the sets of admissible translations and rotations. In D4.2 the feedback control of the head motion by a gradient-ascent strategy has been defined. Therein, given the belief $\hat{p}(x_k|z_{1:k})$, the rigid motion, u_k, applied to the head is defined as being proportional to the direction of the steepest ascent around, that is, $u_k = 0$ of $F_k(u_k) = \log S_{k+1|k}$.

Following the prospects of D4.2, this first result has been extended to the the determination of the optimum admissible finite motion within the cylinder, $\mathcal{T} \times \mathcal{R}$. The optimization problem (4.29–4.31) then writes as

$$(\mathcal{P}) \,:\, u_k^* = (T_y^*, T_z^*, \phi^*) = \arg \max_{(T_y, T_z, \phi) \in (\mathcal{T} \times \mathcal{R})} F_k(T_y, T_z, \phi). \quad (4.34)$$

Thanks to the expansion (4.32), the projected gradient algorithm can be used to solve $(\mathcal{P})$ numerically. It consists in iteratively updating the value of the decision variable, $U = (T_y, T_z, \phi)^T$ obtained through the conventional-gradient-ascent method by projecting it onto the closed convex set $\mathcal{T} \times \mathcal{R}$, by means of the projection opera-

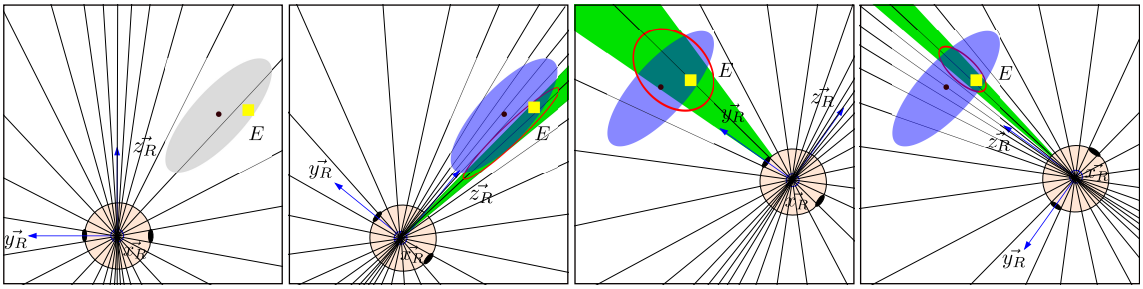


Figure 4.24: Iso-z loci and measurement update for various scenarios. **(a)** Frame $\mathcal{F}_k$ attached to the binaural head **(blue)**, sound-source genuine position **(yellow square)**, confidence ellipse associated to the belief at time k **(grey)**, iso-z_k loci depicting the measurement space **(grey radial lines)**. **(b), (c), (d)** Frame $\mathcal{F}_{k+1}$ **(blue)**, confidence ellipse associated to the next predicted-state pdf at time $k+1$ **(blue)**, iso-z_{k+1} loci **(grey)**, confidence cone associated to the measurement **(green)**, confidence ellipse associated to the next filtered-state pdf – belief at $k+1$ – after the incorporation of z_{k+1} **(red)**

tor

$$\pi_{\mathcal{T}\times\mathcal{R}}(U) \triangleq \arg\min_{x} \{\|U - x\|_2,\ x \in (\mathcal{T}\times\mathcal{R})\}\,. \tag{4.35}$$

This leads to Algorithm 1.

Algorithm 1: Simplified Projected Gradient

Data:

- Moments of the initial belief $\hat{p}(x_k|z_{1:k}) = \mathcal{N}(x_k; \hat{x}_{k|k}, P_{k|k})$ at time k: $\hat{x}_{k|k}, P_{k|k}$
- Maximum admissible translation and rotation of the head: r_{max}, ϕ_{max}
- Step size: γ • Number of iterations: M • Projection operator onto $\mathcal{T}\times\mathcal{R}$: $\pi_{\mathcal{T}\times\mathcal{R}}(T_y, T_z, \phi)$

Output: • $U^* = (T_y^*, T_z^*, \phi^*)^T \triangleq U_M = (T_{yM}, T_{zM}, \phi_M)^T$

$U_0 = [T_{y0}, T_{z0}, \phi_0]^T$
for $i = 0, \ldots, M-1$ **do**
  evaluate $d_i = \nabla F_k(U_i)$, where F_k is defined on the basis of the initial belief $\hat{p}(x_k|z_{1:k})$ at time k;
  set $U_{i+1} = \pi_{\mathcal{T}\times\mathcal{R}}\,(U_i + \gamma d_i)$;
end

Insights into the geometry of the maximization problem, ($\mathcal{P}$) Given a belief $\hat{p}(x_k|z_{1:k}) = \mathcal{N}(x_k; \hat{x}_{k|k}, P_{k|k})$ on the sensor-to-source position at time k, the level sets of the criterion $F_k(T_y, T_z, \phi)$ can be portrayed with regard to the translation and rotation variables T_y, T_z, ϕ. The gradients of $F_k(T_y, T_z, \phi)$ are orthogonal to these surfaces and point to directions of steepest ascent. Restricting to horizontal sections of the admissible cylindrical set, $\mathcal{T}\times\mathcal{R}$, indexed by values ϕ_0 of the rotation variable, leads to contour lines of $F(T_y, T_z, \phi_0)$ with regard to T_y, T_z – see Fig. 4.25. The gradients on these 2–dimensional sections are obtained by setting the third entry of the genuine gradients to 0.

For the forthcoming instances of problem ($\mathcal{P}$), the optimum solution(s) can be observed to lie on the external surface of $\mathcal{T}\times\mathcal{R}$. So, the contour lines of the criterion F_k constrained to its surface, $\partial\big(\mathcal{T}\times\mathcal{R}\big)$, will be also displayed. They will be referenced by the horizontal and vertical angles, α, ϕ, thanks to the bivariate function

$$\tilde{F}_k(\alpha, \phi) = F_k\big(r_{max}\sin(\alpha), r_{max}\cos(\alpha), \phi\big). \tag{4.36}$$

Iso-entropy contour lines for ITD-based exploration Assume that in (4.27), $\bar{l}(\theta_k)$ stands for the Woodworth–Schlosberg approximation of the ITD between two antipodal microphones placed on a spherical head, namely, $\bar{l}(\theta_k) = \frac{a}{c}(\theta_k + \sin(\theta_k))$ for

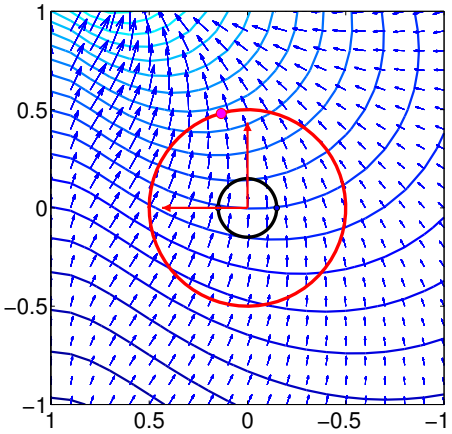
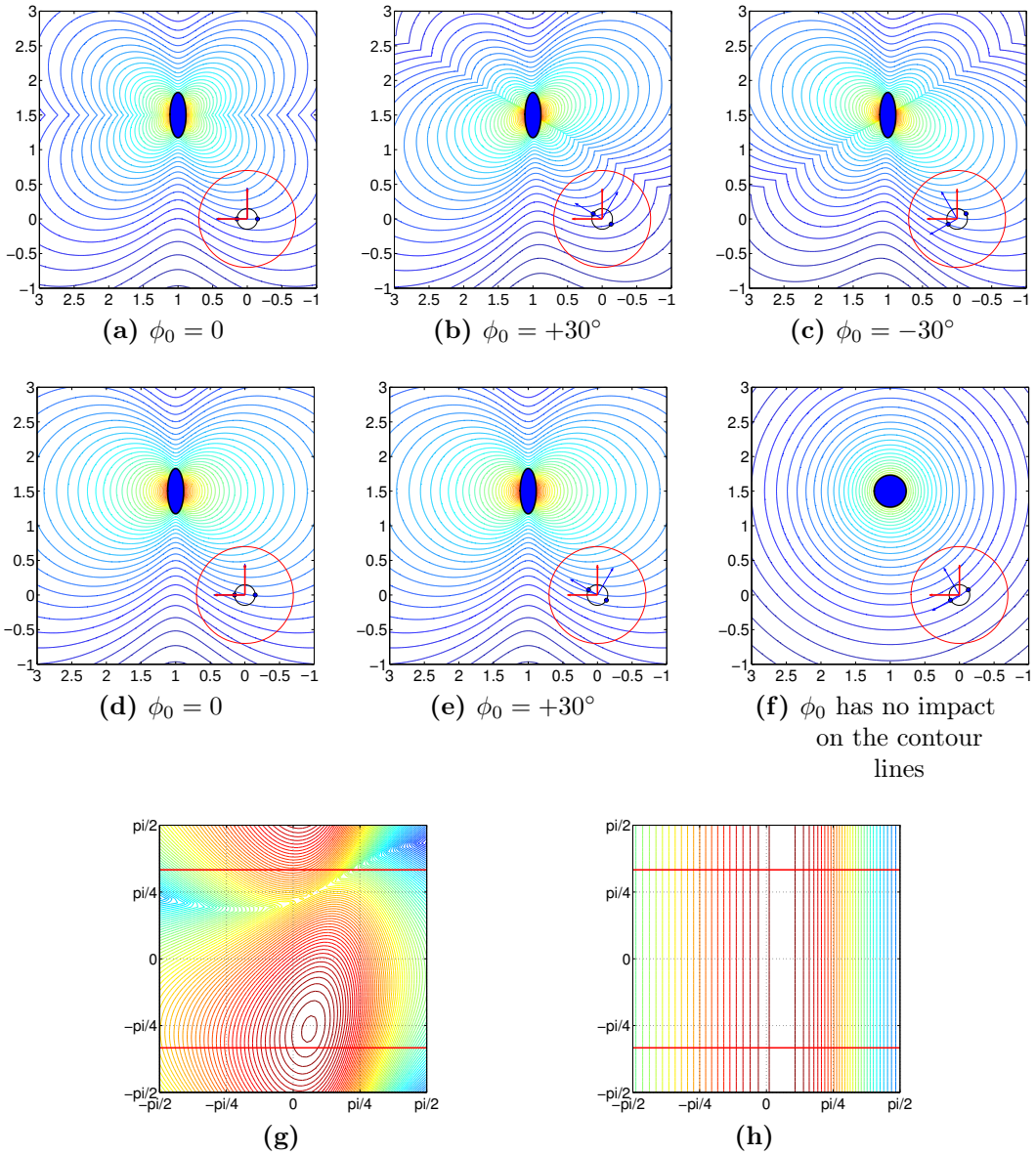


Figure 4.25: Contour lines and local gradient vectors of the criterion $F_k(T_y, T_z, 0)$ with respect to the translation variables, T_y, T_z, that is, when no subsequent rotation is applied to the head – $\phi = \phi_0 = 0$. The **red circle** delimits the admissible translations. The **magenta spot** depicts the constrained local maximum

$|\theta_k| \in [0, \frac{\pi}{2}]$, $\bar{l}(\theta_k) = \frac{a}{c}(\pi - \theta_k + \sin(\theta_k))$ for $\theta_k \in [\frac{\pi}{2}, \pi]$ and $\bar{l}(\theta_k) = \frac{a}{c}(-\pi - \theta_k + \sin(\theta_k))$ for $\theta_k \in [-\pi, -\frac{\pi}{2}]$, with c the velocity of sound [1]. Then, the iso-z_k loci are similar to those depicted in Fig. 4.24. The contour lines of $F_k(T_y, T_z, \phi_0)$ are plotted on Figs. 4.26a–4.26c regarding T_y, T_z for various subsequent rotations, ϕ_0, of the head, given an initial frame, $\mathcal{F}_k$, and a confidence ellipse, $\mathcal{E}_k$, describing the belief $\mathcal{N}(x_k; \hat{x}_{k|k}, P_{k|k})$, where $\hat{x}_{k|k} = (1, 1.5)^T$. The set of admissible translations is displayed, so as to deduce the constrained local maximum on the slice of $\mathcal{T} \times \mathcal{R}$ defined by ϕ_0.

In Fig. 4.26a, the sensor undergoes a pure translation followed by no rotation. The contour lines of the criterion appear to be distorted – that is, the gradient of the criterion is subject to important local variations – whenever the translation is $T = (1, .)^T$ or $T = (., 1.5)^T$. These distortions can be explained by the aforementioned rules of thumb. In such cases, the head must get closer to the source so as to reach a given value of the information criterion, allowing for a neighboring unrestricted translation to be applied.

Subsequent rotations of the head by $\phi_0 = +30°$ or $\phi_0 = -30°$ turn Fig. 4.26a into Figs. 4.26b–4.26c. The contour lines are changed and, consequently, the maximum is restricted to the slice defined by ϕ_0. It is preferable to apply a rotation of $-30°$ rather than $+30°$, because the optimum for $\phi_0 = -30°$ lies on a "warmer" contour line. Noticeably, the first distortions explained in the above paragraph for a null rotation remain, while the second ones are just rotated by ϕ_0. Also, as the step size between the indices of two consecutive contour lines is constant, and as these contour lines are not regularly spaced, the closer the sensor gets to the source, the higher is the increase in the information criterion F_k.



(a) $\phi_0 = 0$ **(b)** $\phi_0 = +30°$ **(c)** $\phi_0 = -30°$

(d) $\phi_0 = 0$ **(e)** $\phi_0 = +30°$ **(f)** ϕ_0 has no impact on the contour lines

(g) **(h)**

Figure 4.26: **(a, b, c, d, e, f)** Contour lines of the criterion $F_k(T_y, T_z, \phi_0)$ with regard to T_y **(abscissa, in meters)** and T_z (ordinate, in meters). **(g, h)** Contour lines of $\tilde{F}_k(\alpha, \phi)$ with regard to α (abscissa, in radians) and ϕ (ordinate, in radians). In **(a, b, c, g)** [resp. **(d, e, f, h)**], the exploration is based on ITD measurements [resp. on ideal azimuth observations]. The sensor frame in the initial position $\mathcal{F}_k = (O, \overrightarrow{x_R}, \overrightarrow{y_R}, \overrightarrow{z_R})$ is plotted in **red**. The initial estimate of the head-to-source position is $\hat{x}_{k|k} = (1, 1.5)^T$. The **blue** ellipse/circle represents the 99%-probability confidence ellipse associated to the initial belief $\mathcal{N}(x_k; \hat{x}_{k|k}, P_{k|k})$. The **red circle** delimits the admissible translation $T \in \mathcal{T}$. The **blue frame** portrays the orientation of $\mathcal{F}_{k+1}$ if a zero translation were applied. The contours are warm [resp. cold] when F_k or, equivalently, $\tilde{F}_k$ has high [resp. low] values. In **(g, h)**, the **horizontal red lines** depict the limits of the admissible head rotation, which have been set to $\pm 60°$

To get some insight on the maximum value of $F_k(T_y, T_z, \phi)$ on the cylindrical surface of the admissible set, the function $\tilde{F}_k(\alpha, \phi)$ has then been evaluated for the same initial belief. It appears that its maximum is located on $\phi^* = -48°$ (Fig. 4.26g).

In some cases, for instance, Fig. 4.27, $(\mathcal{P})$ has several optima.

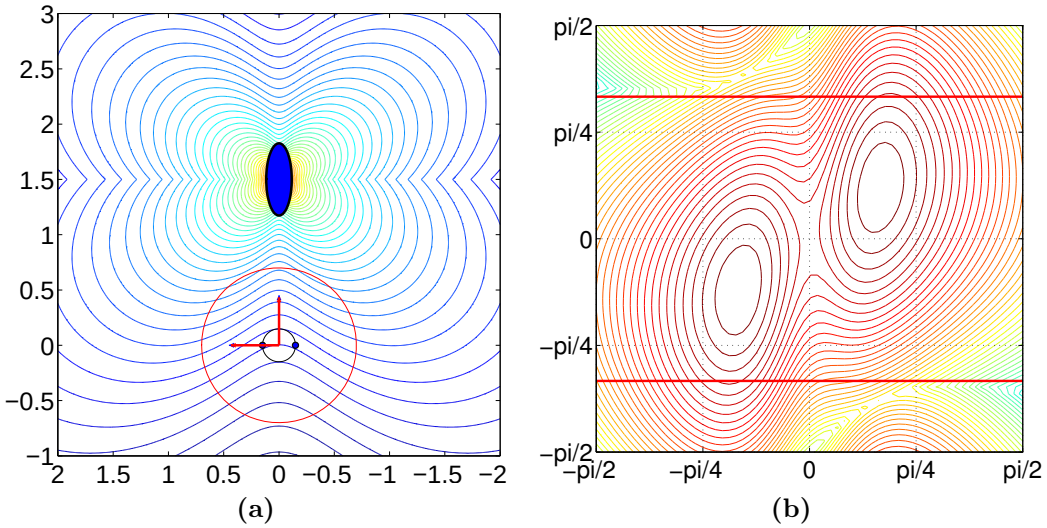


Figure 4.27: For $\hat{x}_{k|k} = (0, 1.5)^T$, **(a)** contour lines of $F_k(T_y, T_z, \phi_0)$ with respect to T_y, T_z **(b)** $\tilde{F}_k(\alpha, \phi)$ regarding α, ϕ when $(\mathcal{P})$ has two solutions. Conventions similar to Figs. 4.26a–4.26h are used

Iso-entropy contour lines for azimuth based exploration This section considers the following observation model.

$$z_k = \theta_k + v_k, \;\; z_k \in \mathbb{R}, \;\; v_k \sim \mathcal{N}(0, R_k). \tag{4.37}$$

Note that observing azimuth measurements contaminated with constant-variance noise is unrealistic in practice. Indeed, when extracting azimuth measurements from the binaural stream, the associated uncertainty is the smaller [resp. all the bigger] as the sound source emits from the front [resp. interaural] axis. Nevertheless, this case has been included because it enables a verification of some intuitive features.

The iso-z loci corresponding to equispaced values of the azimuth measurements are equiangular radial lines passing through O. The confidence cones associated to any measured azimuth then have the same width, they are just rotated images of each other. So, given a belief on the source position evenly spread around its genuine location, the assimilation of such an azimuth measurement intuitively brings the same information if the

sensor remains static or moves on a circle centered on the source, regardless of its orientation.

The analysis of the contour lines of $F_k(T_y, T_z, \phi_0)$ with respect to T_y, T_z shows that they do not depend on the rotation, ϕ_0, (Figs. 4.26d–4.26e). Consequently, the contour lines of $\tilde{F}_k(\alpha, \phi)$ regarding α, ϕ are vertical – Fig. 4.26h. Nonetheless, the contour lines are still distorted for $T = (1, .)^T$ in Figs. 4.26d–4.26e for the same reasons as those explained above. These distortions vanish when the confidence ellipsoid associated to the initial belief is circular – Fig. 4.26f – so that the contour lines become concentric. In this case, the only way to increase the gained information on the source location is to get closer to it, which is in agreement with the above intuition.

Evaluation of the algorithm in simulation with spatialized audio signals The whole three-stage scheme has been implemented on a simulated KEMAR binaural head-and-torso-simulator (HATS) endowed with two translational and one rotational degree of freedom. For the sake of simplicity, the binaural head moves every $T_s = 1\,\mathrm{s}$, then stops in order to acquire binaural signals, extract short-term directional cues – Stage A – and update the belief (4.24) on the source position–Stage B. Its next best position – Stage C– comes from the solution of $(\mathcal{P})$ defined in (4.34) by Algorithm 1 for a Woodworth–Schlosberg measurement equation.

The sound source is static. It emits a non-intermittent white noise filtered by a 1 kHz-bandwidth band-pass filter with 1 kHz central frequency, so as to improve Stage A. It is initialized at the position $X = (1, 2)^T$ in the robot frame $\mathcal{F}_0 = (O, \overrightarrow{x_0}, \overrightarrow{y_0}, \overrightarrow{z_0})$ at time $k = 0$. The admissible movements of the binaural sensor (4.33) are such that $r \leq r_{max} = 0.1\mathrm{m}$ and $|\phi| \leq \phi_{max} = 15°$. Realistic rendering of binaural signals has been simulated in an anechoic environment when the head moves and then listens, by using the TWO!EARS database of Head- Related Impulse Responses (HRIRs) suited to the KEMAR HATS.

Various motions of the sensor have been simulated, namely, the proposed active strategy, a translation along the interaural axis, a circular movement such that the front direction of the head stays tangent to its trajectory, and a random movement – Fig. 4.28a, see page 60. During the five first seconds in all the scenarios, the same rotational movement is applied to the sensor in order to disambiguate front and back, so that at $t = 5\,\mathrm{s}$ the Gaussian-mixture belief can be better approximated by a single Gaussian pdf. The common progress of the audio-motor localization from initial time, $t = 0\,\mathrm{s}$ to $t = 5\,\mathrm{s}$, is displayed on Fig. 4.28c–4.28d. Then, each specific movement is applied from time $t = 6\,\mathrm{s}$ until the end.

It appears that the active motion translates the sensor and rotates its fovea towards the estimated position of the sound source. By computing the Gaussian-moment-matched

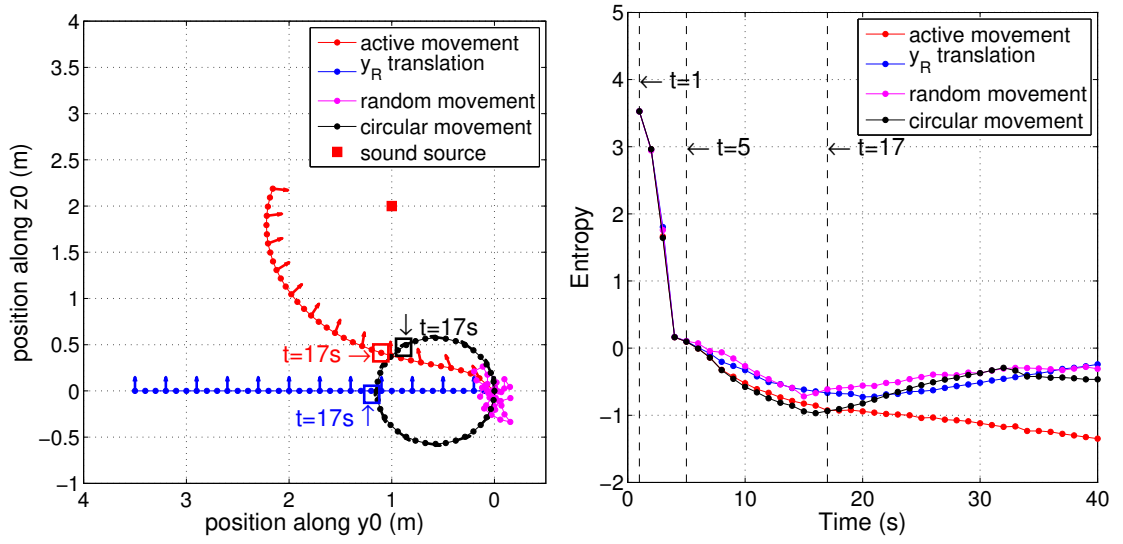
approximation of every state belief, $\sum_{i=1}^{I_k} w_k^i \mathcal{N}(x_k; \hat{x}_{k|k}^i, P_{k|k}^i)$, the entropy, $h(x_k|z_{1:k})$, has been evaluated for the different strategies – Fig. 4.28b. The localization efficiency of the active-motion strategy is clearly the best.

Interestingly, between $t = 5\,\mathrm{s}$ and $t = 17\,\mathrm{s}$, the entropy obtained for the circular motion shows the most important decrease. Though its values are quite the same as for the proposed active strategy at times $t = 6, 7, 8\,\mathrm{s}$, the corresponding beliefs are different. As aforementioned, the active strategy finds the translation and rotation of the head leading to the maximum decrease of the entropy of the posterior-state pdf at the next time step. There is no guarantee that the sequence of N such one-step-ahead optimum motions leads to the maximum entropy decrease at the end of the N steps, that is, constitutes a N-step-ahead optimum motion.

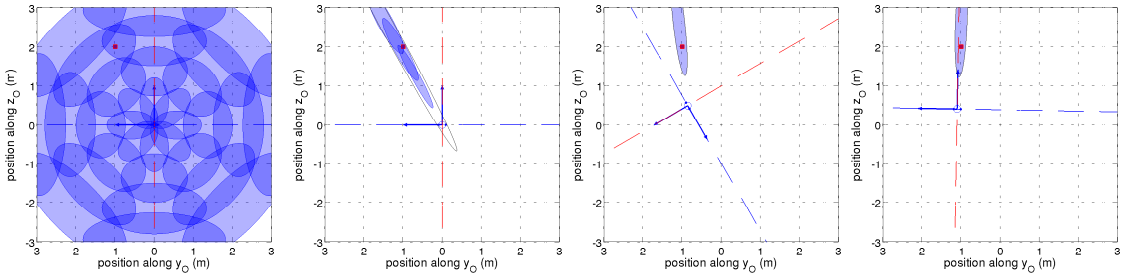
Live experiments on the Two!Ears binaural robot The simulated scenarios have been run on the Two!Ears robot, *Jido*, and its motorized KEMAR HATS, which are endowed with similar degrees of freedom. The results are reported in D5.3 [3].

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: yes



(a) (b)



(c) Initialization (d) Situation at $t = 5\,\mathrm{s}$ (e) Circular movement at $t = 17\,\mathrm{s}$ (f) Active movement at $t = 17\,\mathrm{s}$

Figure 4.28: Simulated sound-source localization for different scenarios. In the circular movement, the front direction is tangent to the circle. The random path is generated by randomly selecting positions on admissible cylindrical sets **(a)** Source position and head trajectories in the world frame – that is, the initial frame $\mathcal{F}_0$. **(b)** Entropy decrease of the posterior state pdf for the various motion strategies. **(c, d, e, f)** Interesting snapshots of the localization process showing the binaural head – front direction in **dashed red**, interaural axis in **dashed blue**, the source in **red**, and the 99%-probability confidence ellipsoids of the hypotheses constituting the Gaussian-mixture belief

4.3 Cognitive-level feedback, e.g., on the basis of labeled Environmental Maps, as built from information taken from the blackboard and from experts (C)

4.3.1 Interpretation of scenes and assigning meaning to their elements (c1)

Head-turning-modulation model

This paragraph first introduces the *Head-Turning-Modulation*(HTM) model , in particular, the motivation behind the creation, development and implementation of such a model, its architecture and how it is embedded in the Two!Ears framework, together with the goals that the model reached throughout the project. Its deep functioning will be further detailed in 4.3.2 & 4.3.6.

The HTM model – or *HeadTurningModulationKS* (HTMKS) as has been implemented as a *Knowledge Source* within the *Blackboard System* [5]– is a low-level-attention module that aims at modulating the head movements of a robot when exploring unknown environment. Modulation of head movements concerns both the action of *triggering* and *inhibiting* them. The *HTMKS* modulates head movements on the basis of two modules, namely,

- **Dynamic-Weighting module** [23, 90] (DWmod) – see Sec. 4.3.2). Its role is to understand the environment being explored by the robot through the notion of *Congruence*. This notion can be described, in the context of the HTM model, as a measurement of how important an object is, given the environment that it is existing in
- **Multimodal-Fusion-&-Inference module (MFImod)** – Sec. 4.3.6, [7]. The role of this module is to learn the relationship between the different modalities that characterize the concept of *object*, namely, in the scope of the Two!Ears project's audio and visual modalities

These two modules lead to the computation of an object that the robot is supposed to focus on. By *focus*, we mean here the spatial area where the robot should turn its head to, that is, where it should concentrate its sensors on in order to gather an optimum amount of information. Thus, motor commands (motor orders) will be triggered by both these modules but for two distinct reasons, that is,

5 The notations "HTM model" and "HTMKS" can be used alternatively. However, we prefer the use of "HTM model" when it comes to denote the concepts which it relies on, and "HTMKS" when it is about its implementation within the Two!Ears framework.

- The motor command triggered by the *DWmod* acts as a low-level attention movement that makes the robot to face source classified as important
- The motor command triggered by the *MFImod* acts as a way to autonomously get some new data in order to feed the artificial neural networks used for the learning phase

One of the main constraints that has driven the creation of the HTMKS is *online real-time learning*. Indeed, in the "learning" community, and especially by those who deal with robotic exploration. Two main paradigms, *robotic listening* and *intelligent systems*, are used to complete such learning tasks, namely,

- Collect **beforehand**, by means of recording, scanning or measuring, optimum data from the environment for creating a comprehensive modelization of the world. These will be provided to the robot system before exploring or evolving this environment
- Acquire **in real-time** the most possible data and, by means of interaction, exploration or reaction, trying to get the information that could be useful for the robot to evolve in the environment

The huge advantage of world-modelizing techniques is that the robot will know its environment perfectly and will be able to behave correctly, quickly and relevantly. However, a main drawback is that it needs the environment to be entirely known beforehand. This is incompatible with Search- &-Rescue (S-&-R) scenarios where, by definition, the environment is assumed to be unknown beforehand.

Nevertheless, *real-time* learning algorithms enable robots to evolve in unknown environments and to learn about it continuously, implying the ability to adapt to unpredictable changes that can occur in realistic scenarios. However, this adaptability goes with the drawback of delaying the instance at which the robot starts behaving adequately. It has thus been decided to conceive a model that provides the robot with the ability to learn in *real-time* with no *a-priori knowledge*, since in S-&-R scenarios the environment has a high probability to be unknown to the robot, thus forcing it to be able to adapt to new situations very quickly.

Regarding such a situation one may consider, as a starting point, a naive robot that is only driven by reflexive behavior, such as turning its head whenever an event[6] pops up in the environment. These *spontaneous head movements* (SHMs) are necessary to gather additional information about surrounding objects. Indeed, for example, listening to a source from a new angle may enhance and refine the localization process.

6 We call every audiovisual object "event" which is acoustically or visually perceivable by an agent, for instance, a person walking or talking, or a glass falling. These "events" become "perceptual objects" when perceived by the robot

SHMs can also lead the robot to face the object, thus adding potentially missing visual information to the system. However, these SHMs are not always needed, especially when the environment is getting more and more well-known to the robot.

Actually, this naive behavior would lead to a lot of head movements, which is potentially hampering the main task in the S-&-R scenario, namely, to identify the putative victims. The idea is thus to put the robot into a state that allows it to accomplish its main task on the basis of the ability of not to be blind/deaf with regard to relevant events occuring in the scene.One can particularly distinguish two situations that would require SHMs

- When the event is *unpredictable*
- When there is an *ambiguity* about the audio or visual label of the event, that is when a modality is missing (event placed behind the robot for instance), or when there are classification errors

The notion of *event predictability* is directly inspired by Shannon work on Information Theory [80]. Within the scope of the HTM model, Shannon's ideas are applied as follows: *The more an event is observed in an environment, the more likely it is to appear again in the future.* Thus, such an event carries less new information than a novel one. Consequently, the overall goal of the HTM model is split into two stages, that is,

- Learn the environment through the notion of multimodal objects and by means of SHMs
- Inhibit these SHMs whenever the system judges that it knows well enough the environment

Figure 4.29 shows the integration of the HTMKS into the Two!Ears Blackboard system, together with a simplified scheme of the internal architecture of the HTMKS. One can also appreciate the scope of the *Head-Turning Modulation model* by considering a so-called *Environmental Map* as resulting from the knowledge created by the HTMKS. Such maps exhibit how the robot perceives and understands its environment.

Model formalization As stated above, the HTM model is implemented as a Knowledge Source (HTMKS). The overall system is encoded in an object-oriented framework organized around three main classes, namely, *RobotInternalRepresentation*,*PerceivedEnvironment* and *PerceivedObject* – see Fig. 4.30. The *RobotInternalRepresentation* class gathers all the information that the robot has observed and processed, such as the different environments it has explored – one *PerceivedEnvironment* instantiation for each of them – and that are populated with all the observed objects – one *PerceivedObject* instantiation for each one. It also contains the audiovisual-categories database created by the environment exploration, as a result of the learning phase performed by *Multimodal-Fusion-and-Inference module* (MFImod). Each time an event is detected by the robot, the

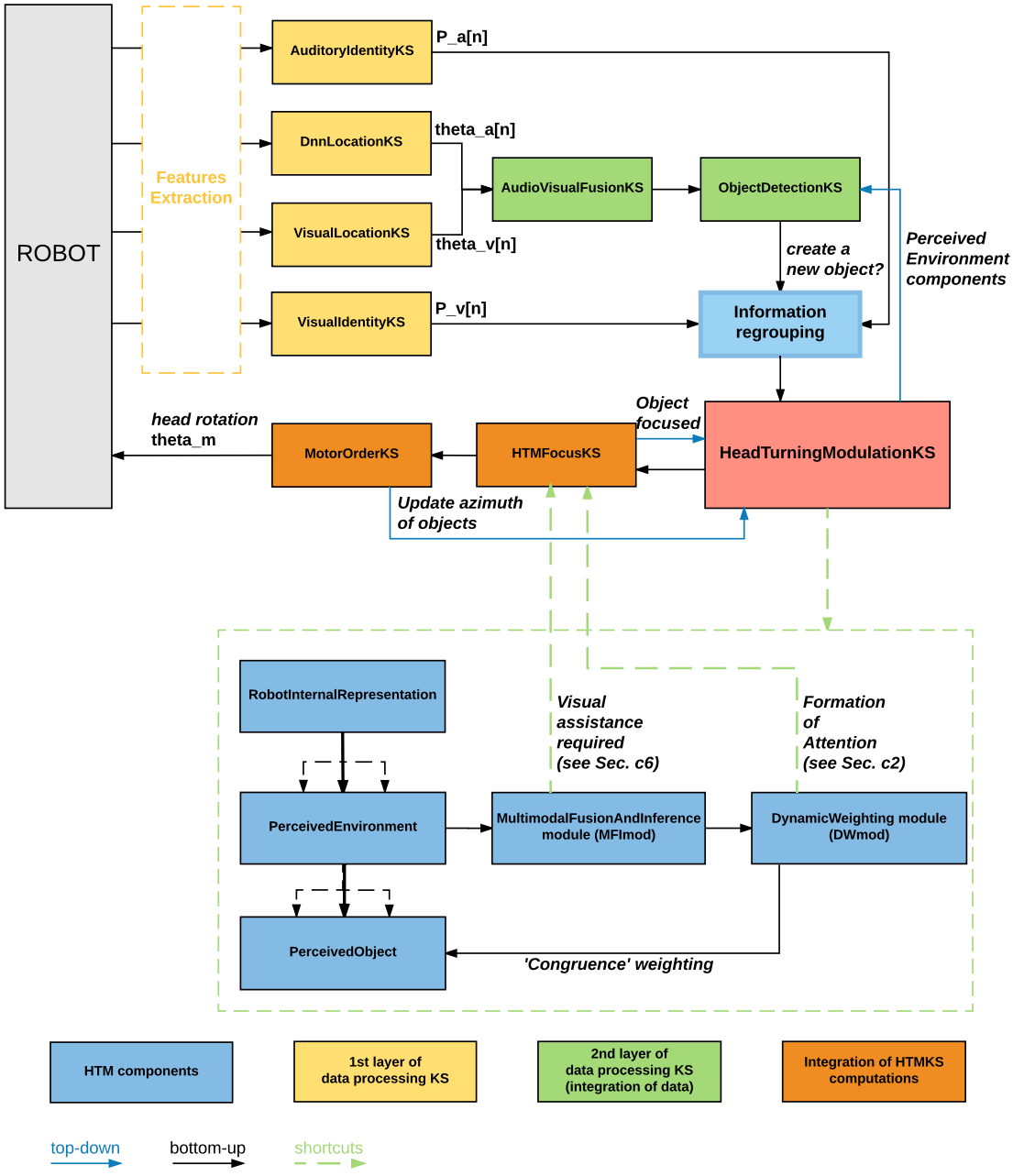


Figure 4.29: Simplified structure of the *HeadTurningModulationKS* and its links to other Two!EarsBlackboard components. **Black arrows** and **blue arrows** denote *bottom-up* (feed-forward) and, respectively, *top-down* feed-back pathways

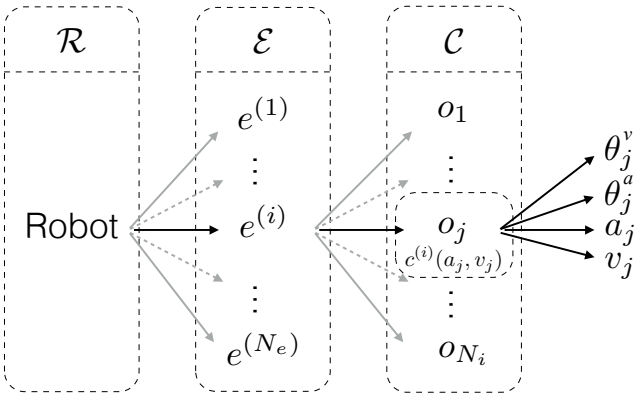


Figure 4.30: Object-oriented paradigm of the HTMKS. A robot is in an environment, $e^{(i)} \in \mathcal{E}$, defined by multiple objects, $o_j \in \mathcal{C}$, characterized by their positions, $\theta_j^{a/v}$, and their audiovisual labels, a_j, v_j – see Sec. 4.3.1

model tries first to assign an audiovisual category to it – this is the role of MFImod – and then computes the congruence on the basis of what the robot has already experienced in the past in the current environment – this is indeed role of DWmod. If it turns out that additional information is required to assign the audiovisual category to the object, a head movement is triggered.

Following this process, if the object has been characterized as *incongruent*, the robot turns its head toward the object. These head movements, as triggered by the two HTMKS modules, are supposed to help the robot to acquire more information about the current event. In the next section, the definitions and notations are detailed which the formulation of the HTM model relies on. They will also be used in Secs. 4.3.2 & 4.3.6.

Definitions and notations Let $\mathcal{R}$ and $\mathcal{E}$ be the robot and environment sets, with

$$\mathcal{E} = \{e^{(1)}, e^{(2)}, \ldots, e^{(N_e)}\}, \tag{4.38}$$

where $e^{(i)} \in \mathcal{E}$ represents the i^{th} environment explored by $\mathcal{R}$, and N_e the number of considered environments. Each environment, $e^{(i)}$, is defined as a set of Objects, o_j, such that

$$e^{(i)} = \{o_1, o_2, \ldots, o_{N_i}\}, \tag{4.39}$$

with N_i the number of detected objects in the environment, $e^{(i)}$. Every object, o_j, is defined by its relative audio and visual angles with respect to the robot, $[\theta_j^a, \theta_j^v]$, and auditory and visual labels, $[a_j, v_j]$, such that

$$o_j = \{\theta_j^a, \theta_j^v, a_j, v_j\}. \tag{4.40}$$

The relative audio angle θ_j^a is provided by the *DnnLocationKS*. The relative visual angle θ_j^v is provided by the *VisualLocationKS*. The multimodal labels a_j and v_j are estimated by *AuditoryIdentityKS* and *VisualIdentityKS*, respectively, and are picked among the predefined collections of labels $\mathcal{A}$ and $\mathcal{V}$, accordingly. For instance, one can have $\mathcal{A} = \{\texttt{speech}, \texttt{alarm}, \texttt{crackling}\ldots\}$ and $\mathcal{V} = \{\textit{female}, \textit{siren}, \textit{fire}, \ldots\}$.

Lets now define the *audiovisual categories*, $c^{(i)}(a, v)$, of the i^{th} environment by

$$c^{(i)}(a, v) = \{o_j \in e^{(i)}, a_j = a, v_j = v\}. \tag{4.41}$$

$c^{(i)}(a, v)$ basically represents the collection of objects sharing the same auditory and visual labels, a and v, respectively. All categories of the i^{th} environment are gathered into a set of categories, $\mathcal{C}^{(i)}$, such that $\mathcal{C}^{(i)} = \{c^{(i)}(a, v)\}$. The Knowledge Sources on which the HTMKS relies on are recalled for self-containedness below.

AuditoryIdentityKS This KS[7] outputs a vector $\mathbf{P}^a$ whose dimension is equal to the number of auditory classification experts available, that is,

$$\mathbf{P}^a[n] = (p_1^a[n], \ldots, p_{N_a}^a[n])^T, \tag{4.42}$$

where p_i^v represents the probability of the current audio frame to belong the the i^th audio/visual category. This probability is provided by each *AuditoryIdentityKS* dedicated to each category that has been individually trained beforehand.

VisualIdentityKS In an identical fashion, this KS[8] outputs a vector $\mathbf{P}^v$ whose dimension is equal to the number of visual categories that can be recognized by the system, namely,

$$\mathbf{P}^v[n] = (p_1^v[n], \ldots, p_{N_v}^v[n])^T, \tag{4.43}$$

Similarly to (4.42), p_i^v represents the probability of the current visual frame to belong the the i^{th} visual category. Because of implementation considerations, and different from the *AuditoryIdentityKS*, there is only one *VisualIdentityKS* that gathers the information about all the visual categories recognizable by the system.

DnnLocationKS This KS[9] outputs a vector, $\mathbf{T}^a$, whose components are the probabilities of a sound source to be located in a given angle within a 360°-wide range by steps of 5°,

7 see D**6.1.3**, §**3.3.3**, and D**3.5**, §**4.2**
8 see D**5.3** & D**6.1.3**
9 see D**6.1.3**, §**3.3.1**, and D**3.5**, §**3.4.1**

that is,

$$\mathbf{T}^a[n] = (\theta_1^a[n], \ldots, \theta_N^a[n])^T, \tag{4.44}$$

where $N = 360/5 = 72$ angles.

VisualLocationKS This KS[10] outputs a vector, $\mathbf{T}^v$, collecting the angles, θ_j^v, of the detected objects within the field of view of the robot. These angles also take into account the azimuthal position of the torso and of the head with respect to the resting position (defined beforehand), so that,

$$\mathbf{T}^v[n] = (\theta_1^v[n], \ldots, \theta_{fov}^v[n])^T + \theta_{head} + \theta_{torso} \mod 360, \tag{4.45}$$

where *fov* denotes the field of view of in tereovision.

Building a time-variant vector The outputs of the four KSs introduced above[11] are then taken to form a time-varying vector, $\mathbf{V}[n]$, defined as follows,

$$\mathbf{V}[n] = (\mathbf{P}[n]^T, \mathbf{T}[n])^T, \tag{4.46}$$

with

$$\mathbf{P}[n] = (\mathbf{P}^a[n]^T, \mathbf{P}^v[n]^T)^T, \tag{4.47}$$

$$\mathbf{T}[n] = (\mathbf{T}^a[n]^T, \mathbf{T}^v[n]^T)^T. \tag{4.48}$$

In the following, the vector $\mathbf{P}[n]$ will be referred to as the n^{th} *audio-visual frame*, and the vector $\mathbf{T}[n]$ will be referred to as the n^{th} *audio-visual location.*

On the basis of the vector $\mathbf{V}[\mathbf{n}]$, the HTMKS is able to provide the Blackboard system with information that can be understood as belonging to the cognitive level of the Two!Ears system. In detail, we have

- A list of the audiovisual objects that have been *perceived* by the system
- Information on how they have been perceived. Can be different from the ground truth!
- The amount of *congruence* given the environment
- A list of the overall audiovisual categories that have been detected

10 see D**5.3** & D**6.1.3**

11 An additional *ObjectDetectionKS* (see D**6.1.3**, Sec. 3) has also been implemented in order to deal with discontinuous signals (such as speech or when an object stops emitting and then starts again). It uses the output of the DnnLocationKS or the VisualLocationKS, depending on what modality is available, and computes if the putative position of the newly detected event matches previous observations of already detected objects

- Information regarding the ability of the system to infer the audiovisual categories correctly on the basis of only one of the two modalities
- A motor command that can be used to turn the head[12]

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on TWO!EARS: no
Runs on the robot: no

Multiple-source localization with a continuously moving head

This task has been realized with *The Bochum Experimental Feedback Testbed* (BEFT) in emulation mode. BEFT is described in Chap. 5. Herein, the virtual robot is prompted to infer the x–y-plane locations of three surrounding sound sources. In-place rotation of the robotic agent is employed to get rid of front/back-ambiguities. The simulation is driven by the TWO!EARS framework, relying on a set of knowledge sources that have been tailored to the given analysis task. For more details, refer to Sec. 5.1.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on TWO!EARS: yes
Runs on the robot: not intended

4.3.2 Formation of attention and attention-based control of feedback processes (c2)

Dynamic-Weighting module Formation of attention and attention-based control of motor feedback processes are the main part of the HTMKS and are handled, in particular, by the *Dynamic Weighting* module (DWmod). Let's recall that the objective of the DWmod is to trigger reflective head movements as a way to drive the attention of the robot towards important targets. This objective is close to the design of *attentional filtering* systems, which have to be designed with the following questions in mind.

12 If the object is localized at an angle exceeding to the turning limitations of the head/neck of the robot, which is 90° on *Jido*, a motor command is also sent to the torso, so that the robot can reach the intended position

- What is an important event?
- How to react with minimal knowledge about the environment?
- How to avoid the curse of dimensionality inherent to full-world-modeling approaches?

The first question received a lot of regard during the last decade. In 2008, Ruesch et al. [78] successfully developed a powerful model of an attentional filter based on the saliency of multimodal inputs. This algorithm provides the robot with the ability to detect what is the important event or object in a restricted environment, but does not take into account the context: *A talking face is usually considered more salient than a silent one* [78]. In the audio community, saliency is also an efficient way to characterize sounds in complex environments: *Salient sounds are defined as those sounds that can be noticed without attention* [31]. From the same authors stems: *It is pre-attentive and deals with sounds that grab a listener's attention.* Auditory saliency is also used in speech processing for instance [51] where very perfomant and innovative models are used to combined sound processing and low-level attention on the basis of the characteristics of the sound such as intensity, frequency contrast, temporal contrast, orientations or pitch distribution.

The DWmod differs from these approaches in the following ways.

- **The acoustic content** is not taken into account. That is, what the DWmod aims at adressing the matter of important audio events not in terms of acoustical properties but rather in terms of higher level information that these events contain
- The DWmod uses **feedback through a top-down approach**, where the experience and the knowledge that the robot gathers while exploring the environment will form the ability to make the notion of *importance* arise

Despite the fact that *saliency* is a fundamental feature of an object with regards to its surroundings, it can not be sufficient to determine whether an event requires the attention of a robot, from the point of view of HTMKS paradigm. Indeed, to use again a case studied by [78], even if a very colorful object is visually detected faster than a pale object, if the environment is full of colorful objects, pale ones thus carry more information than colored ones do.

More recent models take into account contextual features: Nguyen et al. in 2013 [64], and Ivaldi et al. in 2014 [47] have developed a powerful algorithm as a contribution to the cognitive architecture of the MACSi project[13]. Integrated on the iCub platform, their algorithm enables a robot to actively and interactively learn the objects populating the environment by experiencing actions with regard to them. Behaviors relying on *Curiosity* and *Intrinsic Motivation* have thus been modeled to enable the robot to understand its environment in a more subtle and relevant way – see [79] for a proposal of a unified human motivations taxonomy. For instance, [46] defines *Uncertainty Motivation* as the attraction for novel

13 http://macsi.isir.upmc.fr

stimuli. [77] defines *Information-Gain Motivation* as the *Pleasure of learning* that guides the robot to minimize the level of uncertainty of its knowledge of the environment.[22] defines *Empowerment Motivation* as a behavior that encourages the sequence of actions that will lead to the acquisition of the maximal amount of information by the robot's sensors – see [5] for a case study of in 100 ms intrinsically motivated robots.

All the models developed on the basis of motivation show very good results on exploratory robots. However, whereas these models aim at determining a particular area of the environment to go to, or the next more relevant action to be performed in relatively simple environments, the proposed Dynamic Weighting model acts as a low-level attentional filter motivated by *novelty detection*. The ambition of the present work is to provide a simple –*but not simplistic*– algorithm that enables the robot to filter the environment it is experiencing *without a-prior knowledge*.

The notion of Congruence As described above, DWmod aims at providing a robot with the ability to assign importance to objects in the environment it is experiencing. This is achieved by considering object apparition through the prism of *Congruence*. In Algebra, two plane figures are congruent if they share similar features (such as size and shape). From the point of view of Biology, the notion of congruence, and particularly its opposite, *incongruence*, is reflected by an electrical cortical response called *Mismatch Negativity* (MMN, see [63] for a review). MMN occurs in every sensory area of the brain[14] when an odd stimulus arises among other predictible stimuli. MMN occurs at about 100 ms after the onset of the odd stimulus. This quick reaction is considered to be an alert mechanism that leads to a quick behavioral response [4]. By extension of both these mathematical and biological definitions, congruence will thus be defined along

- The features shared by two perceptual objects, such as visual and acoustic labels
- The links that exist between a perceptual event and a given environment

If an object has been detected as incongruent, a quick head movement will be triggered into the direction of the object. This head movement will have several consequences regarding perception. The most striking evidence is that when an audiovisual object emitting sound but is out of sight, a head turn toward the object will be initiated, thereby

- Enhancing the estimated position of the object, by updating its ITD and ILD
- Enhancing discrimination from other sound sources present in thes surroundings – The issue of sound-source separation, as apparent in the Cocktail-party problem)
- Providing the missing visual information as regards the object. This leads to a more accurate perception of the object

14 These are the areas that process sensory information such as audition, vision, smell, touch, and taste

Weights computation In order to decide if an object in the environment is of interest, each object is associated to a weighting function, $w(o_j)$. In all the following, an audio-visual object, o_j, will be classified as incongruent *if other objects belonging to the same category, $c^{(i)}(a_j, v_j)$, have not been detected by the system in the past.* This classification between congruent/incongruent objects will in fact be based on the object-weighting function, $w(o_j)$, with $w(o_j) \in$ [-1;1]. In all of the following, $w(o_j) = -1$ represents a highly congruent object, while $w(o_j) = 1$ indicates a highly incongruent object. Note that the former case will not produce any movement of the robot, while the latter will trigger SHM in the direction of θ_j. First, and based on the previous definitions, lets define the pseudo-probability, that is, the frequency) $p(c^{(i)}(a_j, v_j))$ as

$$p\left(c^{(i)}(a_j, v_j)\right) = \frac{|c^{(i)}(aj, v_j)|}{N_i}, \tag{4.49}$$

with

$$\sum_{n=1}^{|\mathcal{C}^{(i)}|} p(c^{(i)}(a_n, v_n)) = 1, \tag{4.50}$$

where $|.|$ denotes the set cardinality. The pseudo-probability, $p(c^{(i)}(a_j, v_j))$, can be considered as the likeliness of an object, o_j, to belong to category, $c^{(i)}(a_j, v_j)$. On this basis, the weight, $w(o_j)$, of the object can then be defined as

$$w(o_j) = \begin{cases} 1 & \text{if } p(c^{(i)}(a_j, v_j)) < K_i, \\ -1 & \text{else,} \end{cases} \tag{4.51}$$

where K_i denotes a frequency threshold. (4.51) clearly shows the relation between a high object's weight, $w(o_j)$, and a low probability of the object's category occurrence, $p(c^{(i)}(a_j, v_j))$. Thus, if the object appears in the current scene, it will be categorized as *incongruent* and, consequently, a SHM will be triggered. In all of the following, the threshold, K, is set to

$$K_i = \frac{1}{|\mathcal{C}^{(i)}|}, \tag{4.52}$$

that is, $w(o_j) = 1$, which means that the object is incongruent if its probability to belong to a certain category, $p(c^{(i)}(a_j, v_j))$, is smaller than a random choice among equiprobable categories. However, (4.51) exhibits a very naive weighting strategy. The proposed binary decision, while very simple, may indeed lead to inconsistent behavior when dealing with multiple objects in the environment at the same time. Additionally, classification errors may introduce spurious temporary categories, $c^{(i)}(.)$, which could be rated as incongruent although they are not relevant. Time filtering of the decision is thus introduced in the proposed weighting function so as to increase the robustness of the approach. Inspired by the N100 cortical-wave pattern arising at around 100 ms after the onset of an odd stimulus – compare Sec. 4.3.2 – a modified weighting function is proposed. Two smooth

symmetric sigmoid functions in the range of $[-1;1]$ are used to introduce a dynamic weighting, $w(o_j)[n]$, of an object, o_j, with

$$w(o_j)[n] = \begin{cases} 1/(1+100\,e^{-2n}) & \text{if } p(c^{(i)}(a_j, v_j)) < K, \\ 1/(1+0.01\,e^{2n}) - 1 & \text{else,} \end{cases} \tag{4.53}$$

where n represents the time-frame index. $w(o_j)[n]$ is plotted in Figure 4.31 as a function of the time-frame index. For a frame length of $T_w = 20\,\text{ms}$ [15] $w(o_j)[n] \approx 1$ at time

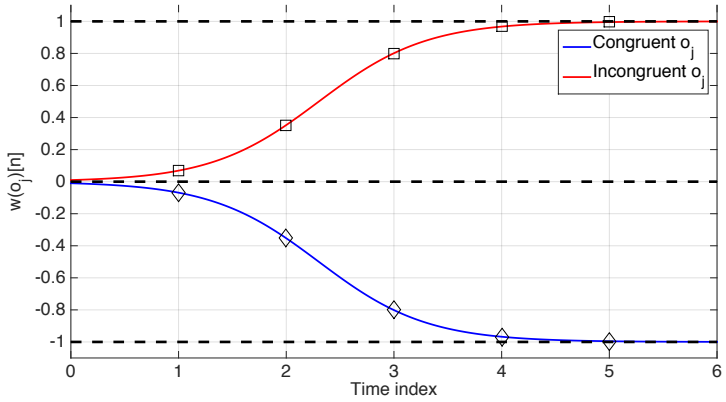


Figure 4.31: Object weight, $w(o_j)[n]$, as a function of time. Depending on the object congruence, one of the two functions is selected. Dots indicates the discrete time steps where values are selected

$t = 100\,\text{ms}$, that is, $w(o_j)[n]$, has been selected to mimic the dynamic of the biological *Mismatch Negativity* phenomenon – see Sec. 4.3.2.

Head-movement decision As a last step, once the weights $w(.)$ of a new object has been computed, one has to decide whether a SHM has to be triggered. A motor command, $\theta_m[n]$, is produced to turn the robot's head to an angle θ_j^a at time index n according to

$$\theta_m[n] = \begin{cases} \theta_j^a & \text{if } w(o_j)[n] > 0.98, \\ \theta_m^a[n-1] & \text{else.} \end{cases} \tag{4.54}$$

A threshold value of 0.98 has been selected based on the weighting-function dynamic of 100ms, – see (4.53).

The resulting pseudo-code of the proposed Dynamic-Weighting model is shown in Alg. 2.

15 This choice will be justified in the next section

Algorithm 2: Pseudo-code of DWmod

loop ▷ % *For each time index*

% *Compute proba. of categ. and objects weights*
forall the $c^{(i)}(a_j, v_j) \in \mathcal{C}^{(i)}$ **do**
Compute $p(c^{(i)}(a_j, v_j))$ according to (4.49)
Compute $w(o_j)$ according to (4.51) or (4.53)
end

% *Find the object associated to the maximal weight*
$[W, idx] = \max_j (w(o_j))$

% *Head turns toward object idx if* $w(o_{idx}) > 0.98$
Compute motor order θ_m according to (4.54).

% *Add current object category in category list*
if $c^{(i)}(a_j, v_j) \notin \mathcal{C}^{(i)}$ **then**
$\mathcal{C}^{(i)} = \{\mathcal{C}^{(i)}, c^{(i)}(a_j, v_j)\}$
end

end loop

Softwarestatus

Data/algorithm available: Yes
Code written and tested: Yes
Implemented on Two!Ears: Yes
Runs on the robot: Yes ongoing

4.3.3 Performing quality judgments from the listener's point of view, based on internal references (c3)

One of the goals of Two!Ears is to investigate how listeners rate the quality of different multi-channel audio presentation systems. In the context of music presentation, it is not the goal of those systems to recreate the exact physical conditions that were present during the music recording, but to create a convincing experience during listening. From a researcher's perspective this leads to the problem, that there is no reference that can be presented to the listeners for comparison of the stimulus under test. The listeners quality judgements are only influenced by their internal references.

This does not only confront auditory modeling of those quality judgments with great challenges. Already at the stage of designing listening experiments great challenges have to be solved. To investigate and solve these we conducted various experiments with the aim of comparing different multi-channel audio systems.

The first task was to create non-trivial musical pieces for the different systems that are comparable and have no direct influence on the quality judgments of the listener. To achieve this, we developed a mixing chain for pop-music that, on the one hand, allowed a high amount of comparability between the mixes for the different systems but, on the other hand, also allowed to use different reproduction systems to their best abilities [43]. Using the developed pipeline we created mixes of four different pop-music pieces for *wave-field synthesis, stereo*, and *5.1 surround*. The finished mixes are available as loudspeaker feeds in the TWO!EARS database (see D 1.3) and are published in [41]. The single signal feeds of the recordings used for creating the mixes are available at [42].

The different reproduction systems were situated in the same studio-like room which was equipped with a 56-channel circular loudspeaker array with a radius of 3 m plus a sub-woofer. For wave-field synthesis all loudspeakers were used, for stereo and 5.1 surround corresponding ones were selected. Using this setup we conducted an experiment comparing the different reproduction systems for the four music mixes by running a paired-comparison test. The listeners had the ability in each case to switch freely between the two presented stimuli. The result showed a clear preference for the reproduction systems applying more loudspeaker for all four pieces of pop-music [44].

In a follow up experiment, we investigated the influence of the mixing process itself on the preference ratings of the listeners. This allows to get insights when the preference rating is dominated by the reproduction system or by the musical mix. In addition, it provides us with data for developing a prediction model, because it shows the influence of single mixing parameters. Actually, with regard to the rated preference, these often correlate with individual perceptual attributes.

For this experiment one song of the first experiment was selected and systematically modified for the wave-field-synthesis system. We selected four mixing parameters, namely, *reverb*, *EQ*, *compression*, and *positioning of foreground elements*.

The second experiment was again performed with the real loudspeaker array. This is problematic, as it is not easy to provide adequate binaural input signals to the model. We thus created the binaural signals by using a binaural simulation of the loudspeaker array. As this simulation is not transparent with regard to all perceptually relevant attributes, it could well be that the simulation itself influences the preference ratings of the listeners. To quantify this potential influence, we repeated the second experiment using dynamic binaural synthesis and presented the stimuli over headphones to the listeners.

We further performed a test on finding the sweet spot. The results revealed that visual feedback information on the actual position of the listener has an influence on the results. This and all other experiments presented in the current section are described in more detail in D 6.2.3, and all results and stimuli are part of a public database – see

D 1.3.

Softwarestatus

Data/algorithm available: yes
Code written and tested: no
Implemented on Two!Ears: no
Runs on the robot: no

4.3.4 Initiating robot maneuvers for scene exploration, for example, for object-distance determination, approaching sources, triangulation (c4)

Scenario [A] Demonstrate at least three sources (fire, siren, human). Find and "save" the human by approaching the corresponding source

This task is naturally incorporated in the task defined in the below paragraph below, that is "Building of Environmental Maps via triangulation". For more details, also refer to Secs. 5.2, 5.3, and 5.4.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: not intended

Building of Environmental Maps via triangulation

This task has been realized by running *The Bochum Experimental Feedback Testbed* (BEFT) in emulation mode. BEFT is described in Chap. 5. The virtual robot employs active exploration to infer the x/y–plane positionss of multiple surrounding sound sources. Task planning is integrated to enable reasonable rescue behavior of the robotic agent. The simulation is driven by the Two!Ears framework, relying on a set of knowledge sources that have been tailored to the given analysis task. For more details, refer to the Sects 5.2, 5.3, and 5.4. Note that this task naturally envelopes the task "Scenario [A]", that is, at least three sources (fire, siren, human). Finding and "saving" the human by approaching the corresponding source is demonstrated as defined in the paragraph above.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes

Implemented on Two!Ears: yes
Runs on the robot: not intended

4.3.5 Keyword spotting (c5)

For keyword recognition, experiments have been performed with features derived according to the Two!Ears front-end and with the *Kaldi* speech-recognition toolkit [72]. Three different feature-streams were considered, including acoustic-only recognition, lip-reading, and audio-visual keyword recognition. For details of the fundamentals of audio-visual speech recognition, see D3.5.

Currently, in order to address the difficult task of optimally integrating the different feature streams mentioned above, we have introduced a state-based integration scheme that uses dynamic stream weights in DNN-based audio-visual speech recognition. The dynamic weights are based on a time-variant reliability estimate that is derived from the acoustic input. We show that this state-based integration is superior to a simple concatenation of audio-visual features. The suggested dynamic weighting mechanism is even able to outperform a fixed weighting approach based on oracle knowledge of the true signal-to-noise ratio.

Keyword recognition will be used as a high-level cognitive module, for example, to recognize alarm situations and, hence, trigger appropriate S-&-R behaviour of the robot. In addition, the phonetic state, which is implicitly available within keyword recognition, is useful to improve source-segregation algorithm – see [97].

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: not yet

4.3.6 Requesting visual assistance through visual-object localization and identification (c6)

Multimodal Fusion-&-Inference module The visual modality is of high interest within the scope of the *Head Turning Modulation* model. Indeed, it is necessary to enable the robot to create its own internal representation of the explored environment through multimodal objects. Within the HTM model, the *Multimodal Fusion-&-Inference* module (MFImod) constitutes the part where the notion of *audiovisual objects* arises in the robot through online learning of the events populating the explored environment. The main idea is to *learn the relationship between all the modalities that define an event*, that is, audio and

vision. Learning this relationship is useful for two reasons.

- Understanding the environment through an object-based comprehension
- Being able to infer a missing modality and still have access to its putative labels, for instance, when the object is placed behind the robot

Since the HTM components (DWmod & MFImod) are designed to be online learning algorithms, they currently rely on localization and recognition outputs – see Sec. 4.3.1. However, these algorithms inevitably exhibit some errors implying that it is not possible to rely on the current information that they put out. It is thus necessary to take these errors into account and correct them. This is also one of the goals of the MFImod, and an essential one to form relevant audiovisual categories. Visual assistance is then required for two distinct processes, that is,

- Linking the audio data, localization and/or identity, with visual data
- Gather some additional data as a feedback used to confirm the quality of the inference made by the MFI module

As shown in Fig. 4.29, the MFImod is placed just before the DWmod, meaning that this module is thought to feed the DWmod with the data that it has processed. Indeed, in order to compute correctly the *congruence* of an audiovisual object – see Sec. 4.3.2 – it is necessary to have correct and reliable audiovisual-category classification. The MFImod has two main goals that are handled by the learning algorithm which it is based on.

- Learning the link between the audio & visual modalities
- Correcting the errors that can occur from the KSs that it relies on

An audio-visual category, $\mathcal{C}$, of an audiovisual frame is defined as the concatenation of the ground-truth classification of the current audio and visual data, with respect to their position in the environment. Let's consider, for instance, $N_a = 3$ auditory identification KS modeling the following sound categories: $\mathcal{C}_1^a = \{\texttt{speech}\}$, $\mathcal{C}_2^a = \{\texttt{knock}\}$ and $\mathcal{C}_3^a = \{\texttt{alert}\}$. In the same vein, let's imagine $N_v = 2$ visual identification KSs that model the visual categories $\mathcal{C}_1^v = \{door\}$ and $\mathcal{C}_2^v = \{female\}$. Then, if a female is speaking in front of the robot, the audio-visual category output by the MFI, $\widehat{\mathcal{C}}$, is expected to match the real category, $\mathcal{C} = \{\mathcal{C}_1^a, \mathcal{C}_2^v\}$. Moreover, the MFImod output is expected to match the real category even if either (a) the audio and/or visual KSs produce wrong classification results, or (b) audio or visual data are missing. Figure 4.32 shows the MFI internal structure that exhibits two parts.

- The categorization of the n^{th} audiovisual frame[16] computed by a Multimodal Self-

16 Recall that it has has to be keept in mind that the MFI performs active data inference *on the basis of the classifier outputs*, and are not based on the audio or visual cues extracted from the raw signals

Organizing Map (M-SOM)

- The motor command triggered by this M-SOM to confirm the missing data inference

Thus, MFImod can be understood as an *active* classifier-fusion system that estimates the audio-visual category, $\widehat{\mathcal{C}}[n]$, of a perceived object. The M-SOM is directly based on a traditional SOM, which is an artificial neural network that provides a discretized representation of an input space in an unsupervised way [53]. In other words, a classical SOM provides a way to visualize high-dimensional data through a low-dimension projection, preserving the input data topology. In practice, a 2–D map is often used to represent the input data through a 2–D arrangement of nodes, r_{ij}, each of them being associated with (a) a weight vector $\mathbf{w}_{ij}$ of the same dimension as the input data vectors and, (b) a position, (i, j), in the map space.

The learning phase aims at spatially organizing the map in such a way that every node, r_{ij}, represents a particular point of the input space. In our case, this leads to a 2–D map with dimensions $N_a \times N_v$ in which every node r_{ij} is associated with a weight, $\mathbf{w}_{ij} = (\mathbf{w}_{ij}^{a\,T}, \mathbf{w}_{ij}^{v\,T})^T$ – with $\mathbf{w}_{ij}^a$ and $\mathbf{w}_{ij}^a$ of size N_a and N_v, respectively – and represents a specific distribution of audiovisual probabilities, p_i^a, *and* p_j^v – in other words, an audiovisual category. Learning such a map traditionally requires two steps.

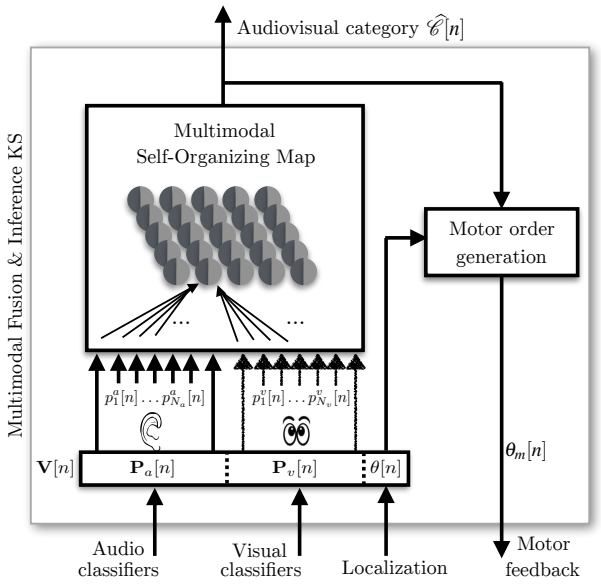


Figure 4.32: Global architecture of the Multimodal Fusion-&-Inference model.

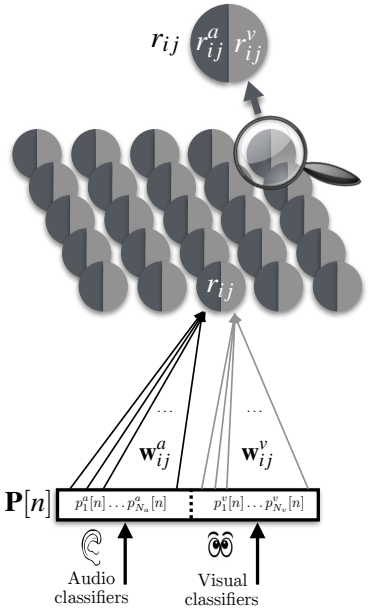


Figure 4.33: Structure of the M-SOM. Each node, r_{ij}, carries two weight vectors, $\mathbf{w}_{ij}^a$ and $\mathbf{w}_{ij}^v$

- Detection of the Best-Matching Unit (BMU), that is, the node, $r_{\text{BMU}}[t]$, whose associated weight vector is most similar to the input vector, being $\mathbf{P}[n]$ learned at iteration, t, namely,

$$r_{\text{BMU}}[t] = r_{IJ}[t], \text{ with } (I,J) = \arg\min_{(i,j)}\{\|\mathbf{P}[n] - \mathbf{w}_{ij}[t]\|\} \tag{4.55}$$

  where $\|.\|$ represents the Euclidean distance, and $(i,j) \in [1,\ldots,N_a] \times [1,\ldots,N_v]$
- Weight adaptation, including a neighborhood function, h_{ij}, which allows for the modification of the input topology around the BMU, as follows.

$$\mathbf{w}_{ij}[t+1] = \mathbf{w}_{ij}[t] + \alpha[t]\ h_{ij}[t]\ \|\mathbf{P}[n] - \mathbf{w}_{ij}[t]\|, \tag{4.56}$$

  with h being defined as a Gaussian neighborhood function, with

$$h_{ij}[t] = \exp\left(-\frac{\|r_{\text{BMU}}[t] - r_{ij}\|^2}{2\sigma[t]^2}\right). \tag{4.57}$$

Once the learning phase is over, the SOM can be used for clustering. Given an input vector, $\mathbf{P}[n]$, the BMU is first localized in the map by using ((4.55)).

Its corresponding weight vector, $\mathbf{w}_{\text{BMU}} = \mathbf{w}_{IJ}$, then carries information about both the audio and visual modalities since $\mathbf{w}_{\text{BMU}} = ({\mathbf{w}^a_{\text{BMU}}}^T, {\mathbf{w}^v_{\text{BMU}}}^T)^T$, with $\mathbf{w}^a_{\text{BMU}} = (w^a_1,\ldots,w^a_{N_a})^T$ and $\mathbf{w}^v_{\text{BMU}} = (w^v_1,\ldots,w^v_{N_v})^T$. Then, the audiovisual category, $\widehat{\mathcal{C}}[n]$, of the input, $\mathbf{P}[n]$, is estimated along

$$\begin{aligned} &\widehat{\mathcal{C}}[n] = (\widehat{\mathcal{C}}^a_A[n], \widehat{\mathcal{C}}^v_V[n]), \text{ with} \\ &\quad A = \arg\max_k w^a_k \text{ ,and } V = \arg\max_l w^v_l. \end{aligned} \tag{4.58}$$

Although a SOM has proven to be a very efficient way to compact and represent high-dimensional data, this tool is not able to cope with missing data. Let's envisage the case where audio or visual data is not available. The corresponding data vector, $\mathbf{P}[n]$, is then made of N_a audio probabilities or N_v visual probabilities set to zero. Such a vector would not make any sense in relation to the data representation in the input space and, thus, leading to the creation of a distinct node during the learning phase. No generalization would then be possible between complete input vectors and vectors with missing data. Consequntly, the traditional SOM architecture must be revisited to cope with the missing data issue.

The Multimodal Self-Organizing Map (M-SOM) Unlike the traditional SOM, the proposed *Multimodal-Self-Organizing Map* (M-SOM) uses *two weight vectors* per node, r_{ij},

namely $\mathbf{w}_{ij}^a$ and $\mathbf{w}_{ij}^v$. Each weight vector is dedicated to a given modality. This specific SOM architecture can also be seen as two distinct traditional SOMs, one dedicated to the audio modality with its own r_{ij}^a nodes, and the second one dedicated to the visual modality with its r_{ij}^v nodes. The two maps are then fused under the constraint $r_{ij}^a = r_{ij}^v = r_{ij}$. The traditional approach allowing for the learning of the weight vectors must be revisited accordingly.

If all modalities are available In such a case, the system will be able to learn the relationship between the audio and visual components, but also to possibly correct wrong classification outputs. Suitable learning and category estimation are described below.

Learning step An *audio-visual* BMU r_{BMU}^{av} is defined along

$$\begin{aligned} &r_{\text{BMU}}^{av} = r_{IJ}[t], \text{ where} \\ &(I,J) = \arg\min_{i,j} \left(\|\mathbf{P}^a[n] - \mathbf{w}_{ij}^a\| \|\mathbf{P}^v[n] - \mathbf{w}_{ij}^v\| \right), \end{aligned} \tag{4.59}$$

which means that the BMU is now defined as being the node whose associated weight vector, $\mathbf{w}_{\text{BMU}}^{av} = ({\mathbf{w}_{\text{BMU}}^a}^T, {\mathbf{w}_{\text{BMU}}^v}^T)^T$, is made of audio *and* visual vectors which are the most similar to the vectors $\mathbf{P}^a[n]$ and $\mathbf{P}^v[n]$, respectively. Once this multimodal BMU is found, the rest of the learning algorithm remains the same and follows Eq. (4.56) and (4.57).

Category estimation Following the same line as the traditional SOM, which could be used to provide a category from the detection of its BMU, one could propose to exploit ((4.58)) on the two audio and visual maps independently. This would result on two – possibly distinct!– BMUs, r_{BMU}^a and r_{BMU}^v, while the two maps have been learned under the constraint $r_{ij}^a = r_{ij}^v = r_{ij}$. Such an approach would t, not take any benefit from the underlying link between the two modalities. Instead, an audiovisual category can still be estimated thanks to ((4.58)), but by using the components w_i^a and w_i^v of the weight vector of the audio-visual BMU $\mathbf{w}_{\text{BMU}}^{av}$ found with ((4.59)). This will result in an estimated category, $\widehat{\mathcal{C}}^{(\text{all})}[n]$.[17]. Importantly, even if some classification errors occur the M-SOM should be able to correct these errors.

If one modality is missing In such a case there is no learning phase. Instead, the current state of the M-SOM is used to *infer* the missing data. Let's consider, as an example, the case where visual data is not available..

17 $^{(\text{all})}$ indicates that both audio and visual data was available

- Audition alone is used to derive the audio BMU r^a_{BMU} in the audio map, whose associate weight, $\mathbf{w}^a_{\text{BMU}}$, can be used to decide on the audio category, $\widehat{\mathcal{C}}^a_A[n]$, with $A = \arg\max_k w^a_k$
- The visual BMU is directly derived from the audio one with $r^v_{\text{BMU}} = r^a_{\text{BMU}}$. Note that this is the step where the link between audio and visual data built during the learning step is exploited!
- Then, the weight, $\mathbf{w}^v_{\text{BMU}}$, associated with the visual BMU, r^v_{BMU}, can be used to decide on the visual category, $\widehat{\mathcal{C}}^v_V[n]$, with $V = \arg\max_l w^v_l$

In the end, the system is able to provide an estimated audio-visual category, $\widehat{\mathcal{C}}^{(\text{miss})}[n] = (\widehat{\mathcal{C}}^a_A[n], \widehat{\mathcal{C}}^v_V[n])$[18], even if no visual data is available. A reciprocal approach can be used when audio data is missing.

Motor-command generation As outlined above, no learning phase occurs if a modality is missing while inference is made to estimate the category of the current object. Taking advantage of having a mobile robot, a motor action could allow to *actively* acquire the missing data. For instance, if the robot is not facing the person currently speaking, it can then turn the head into the person's direction. This would allow the system to grab the visual data that was missing in the first place. This supplemental data can then be incorporated into the learning of the M-SOM, but also be exploited to compare the missing-modality inference (*a-priori* estimation) with the real one obtained after the movement (*a-posteriori* estimation). One important outcome of such an active behavior is that triggering a movement might become less necessary once the M-SOM is able to effectively estimate the audio-visual category. The inference is then considered as reliable enough to inhibit the motor action – unless new audio-visual objects appear in the robot's environment. Let's consider the Kronecker delta $\delta^{(k)}_{ij}[n]$ defined along

$$\delta^{(k)}_{ij}[n] = \begin{cases} 1 \text{ if } \widehat{\mathcal{C}}^{(k)}[n] = (\mathcal{C}^a_i[n], \mathcal{C}^v_j[n]), \\ 0 \text{ else,} \end{cases} \tag{4.60}$$

where $k = \{\text{all}, \text{miss}\}$ denotes if thesound positio category has been obtained without or with missing data[19]. One can then define the *inference ratio*, $q_{ij}[n]$, of the audio-visual category, $(\mathcal{C}^a_i, \mathcal{C}^v_j)$, with

$$q_{ij}[n] = \frac{\sum_{k=1}^{n} \delta^{(\text{miss})}_{ij}[k-1]\, \delta^{(\text{all})}_{ij}[k]}{\sum_{k=1}^{n} \delta^{(\text{miss})}_{ij}[k]}. \tag{4.61}$$

18 where $^{(\text{miss})}$ indicates that there was a missing modality

19 i.e. the exponent (k) indicates whether the category has been obtained without or with missing data, respectively

q_{ij} captures the ratio between the number of confirmed inferences – those for which the inferred category at time $k-1$ has been confirmed *after a head movement* at time k with both audio and visual data available – and the number of times an inference has been made through the M-SOM for a given audio-visual category. On this basis, a head-motor command, $\theta_m[n]$, is generated according to

$$\theta_m[n] = \begin{cases} \theta[n] \text{ if } K_{\text{head}} \leq q_{ij}[n] < 1, \\ \theta_m[n-1] \text{ else.} \end{cases} \tag{4.62}$$

The angle $\theta[n]$, given by the DnnLocationKS, is exploited to turn the head towards the estimated sound-source position at time n. $K_{\text{head}} \in [0, 1]$ and in (4.62) represents a threshold allowing to tune the active behavior. A low threshold will make the system trust quickly in the inferences – and thus will inhibit the head movements – while a high threshold will trigger a lot of head movements as to verify repeatedly whether the inferred audio-visual category is correct. For instance, in a S-&-R scenario, the number of head movements can be set to be lower than in cases where the robot has no time constraints for fully exploring the environment.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: Implementation ongoing

Scenario [B] Demonstrate a scene with at least three sources (2x non-victim, 1 victim). Find the victim by using a combination of auditory and visual cues.

This task has been realized by running *The Bochum Experimental Feedback Testbed* (BEFT) in emulation mode. BEFT is described in Chap. 5. The virtual robot employs active exploration to infer the x/y–plane locations of three sound sources. After acoustic localization, the robotic agent activates its camera system in order to visually discriminate between helpless and physically integer victims. For details, refer to Sec. 5.5.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on Two!Ears: yes
Runs on the robot: not intended

Effects of illumination variations in audio-visual scenarios

This task has been realized by *The Bochum Experimental Feedback Testbed* (BEFT) in emulation mode. BEFT is described in Chap. 5. The virtual robot traverses a building with significantly varying illumination conditions. A virtual camera is attached to the robotic agent, enabling extraction of video footage from a given scenario. The reliability of visual input is then inferred by the machine, using brightness cues derived from the input images. For details, refer to Sec. 5.6.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on TWO!EARS: yes
Runs on the robot: not intended

5 The Experimental Feedback Testbed

As stated in deliverable D4.1 and in the project proposal, "TWO!EARS is meant to challenge current thinking in auditory modeling by replacing common paradigms in this field with a systemic approach, whereby human listeners are regarded as multi-modal agents that develop their concept of the world by exploratory interaction." An *expert-system architecture* (D4.1) drives this behaviour by allowing for cognitive feedback loops and multi-modal cue processing.

Among other fields of interest,TWO!EARS delves into dynamic auditory scene analysis (DASA) in catastrophe settings, using a robot that evacuates perceived victims from the given scene. However, replicating dangerous environmental conditions for experimental tests is challenging in the real world, as these conditions may endanger human subjects or the robotic front-end – see D4.1, Sec. 4.1.2.

To that end, the *Bochum Experimental Feedback Testbed* (BEFT) – as proposed in D4.1, Sec. 1.2.1 – allows to set up experimental DASA scenarios in a virtual environment. Herein, a virtual robot exists that is driven by a *Blackboard system* and retrieves information from the synthetic scene. The*Blackboard system* is a central element of TWO!EARS – refer to D3.1, Sec. 3.1.4 for details. The scenario is rendered within the *Blender 3D* visualization system [12], which enables physics simulation via the *Bullet* [20] physics engine.

Note that in the first version of BEFT, the maneuverability of the robot had been restricted to rotation about the z–axis, thus preventing it from full-scale active exploration of the given scenario. Deliverable D4.2, Sec. 2.1 describes a variant of BEFT, the *Lean Virtual Test Environment* (LVTE). LVTE incorporates a MATLAB^{TM}-based visualization of the virtual scene and allows for translation of the robot in the x/y–plane. With LVTE, first tests in full-scale active exploration and cognition have successfully been performed. Actually, LVTE serves well for baseline testing in active exploration, as shown in D4.2, Secs. 2.2 and 2.3. Yet, compared to the fully-fledged BEFT system, LVTE trades versatility, visual complexity, and multi-modal capabilities for cross-platform compatibility and ease of use –see D4.2, Sec. 2.1.

In the meantime, BEFT has partially been re-designed to allow for full-scale active exploration in a virtual environment, and thus overcome its previous limitations. To ensure full integration with the TWO!EARS system, the current BEFT inherits from the LVTE

by employing a modified *UpdateEnvironmentKS* (D4.2, Sec. 2.2.1) that enables seamless communication with the TWO!EARS Blackboard architecture. Consequently, all cognitive and auditory processing now takes place within the TWO!EARS framework. BEFT, among other things, provides a graphical front-end for virtual handling of dangerous scenarios.

In addition, BEFT retains LVTE's capabilities of low-level auditory processing by providing a specific *SSRInterface* that enables scenario auralization. In these scenarios, the *Sound Scape Renderer* (SSR) [38] generates ear signals for the virtual robot by relying on scene information provided by BEFT – such as sound-source positions and emitted sound signals. The *AuditoryFrontEndKS* (see D3.2, Sec. 3.3.2) then extracts auditory features from the ear signals and makes those features available within the scenario-related Blackboard architecture. Figure 5.1 visualizes the interplay of BEFT and TWO!EARS.

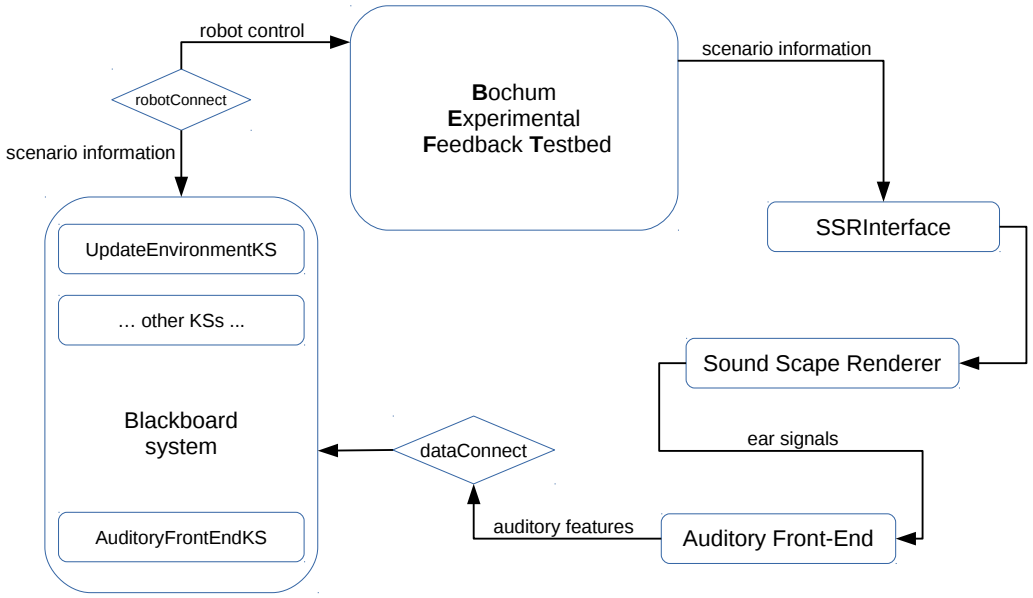


Figure 5.1: Interplay of BEFT and the TWO!EARS system in a virtual audio-visual scenario. Arrows indicate the flow of information and symbolize issued control commands

5.1 Localizing multiple sound sources in BEFT

To qualitatively assess the capabilities of BEFT when analysing audio-visual scenes, a baseline active exploration task was set up. In this task, the virtual robot is placed in an artificial proxy of the ADREAM robot laboratory in Toulouse [65]. Wall reflections are not considered to save on computation time in the auralization process. Three sound sources emit signals that include a human voice yelling for help, a siren, and a barking dog. Each source displays intermitting activation, with artificial silence intervals of 0.5 s in addition to the natural silence intervals contained in each of the sound signals anyhow. In this scenario, the virtual robot has to infer the azimuths of the three sound sources. The Blackboard architecture employed to solve this task is show in Fig. 5.2. Thereby, the *BaselinePlanningKS* represents a simplified variant of the *PlanningKS* discussed in Sec. 5.3 and controls the motion behavior of the robotic agent.

In a first localization attempt, the robot remains in a fixed position, listens for the emitted sounds, and infers the azimuths of all active sources by feeding data from the *Auditory Front End* (AFE) of the TWO!EARS system into the *DnnLocationKS* [57]. This knowledge source, in turn, generates a discrete distribution, $p_f(\Theta)$, that yields the probability of source presence over head-centric azimuth, Θ, in each simulation frame, f.

Using the robot's current head orientation as provided by BEFT, the *VisualizeDNNResultsKS* transforms, $p_f(\Theta)$, into world coordinates, yielding $p_f^w(\Theta)$, and maintains a *source location accumulator*as follows

$$p_{acc}^w = \frac{1}{F_c} \sum_{i=1}^{F_c} p_f^w(\Theta)\,, \tag{5.1}$$

which shows up as peaks growing at the most likely positions of the simulated sources, given all available auditory information up to the current frame, F_c.

Note that summing up probability distributions seems an awkward practice at a first glance, and renders as result of (5.1) a non-probabilistic measure. However, following the probabilistic standard procedure of multiplying all $p_f^w(\Theta)$ would produce severe issues in cases that a sound source mutes, or in cases where source localization for frame f fails. In such situations, the probability peaks for one or more sound source(s) are instantaneously annihilated, and cannot be recovered later.

To circumvent these issues, the summation approach is an alternative that proved well in practice, and was thus taken as the method of choice in the current *proof-of-concept* experiment. Nevertheless, in future expansion stages of the TWO!EARS system, *latent variables* (see [8]) and highly precise localization methods could be introduced to enable

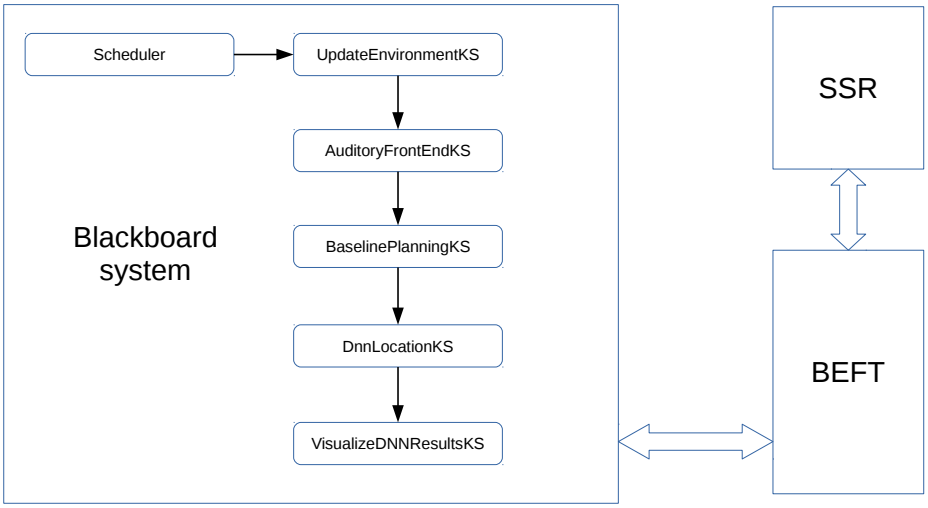
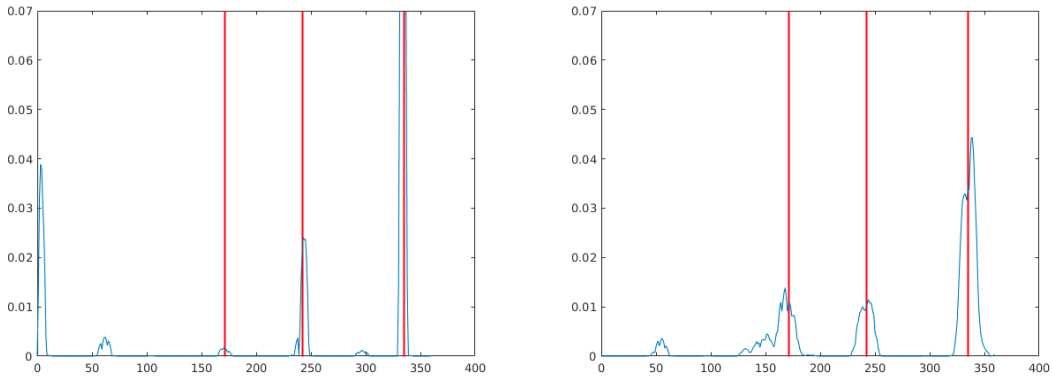


Figure 5.2: Schematic of the Blackboard architecture employed to perform multi-source localization based on the *DnnLocationKS*. **Blueish arrows** indicate the flow of information and symbolize issued control commands. **Black arrows** illustrate the activation sequences of the depicted knowledge sources



(a) **(b)**

Figure 5.3: Results of source-position estimation based on DNN techniques. **(a)** without head rotation. **(b)** using head rotation with an angular velocity of 15°/s. The accumulator content is depicted in **blue**, the true source positions are sketched as **red lines**

a fully probabilistic approach to the given task.

Figure 5.3a shows p_{acc}^{w} for a robot in fixed position after approximately 12 s of simulation time. The red lines indicate the true (physical) source positions, the blue curve represents the content of the location accumulator. One source at $\Theta \approx 170°$ has been missed out. Instead a "ghost" source has been detected, indicated by the false peak at $\approx 3°$. Note that this ghost evolved due to an unresolved front-back confusion in the localization process.

After releasing the fixation of the virtual robot, the device is allowed to turn around its z–axis, with an angular velocity of 15° per second. The predefined panning motion follows the pattern $[-45° \rightarrow +90° \rightarrow -45°]$. Learning from [9], horizontal sweeping should allow the system to perform disambiguation with respect to front-back confusions. Fig. 5.3b displays p_{acc}^{w} for the head sweep approach. The ghost at $3degree$ has now vanished and all sound sources have been localized correctly, as indicated by the prominent accumulator peaks close to the true source positions (red lines).

This basic experiment qualitatively demonstrates the capabilities of BEFT when handling the low-to-mid-level auditory cues like *ITDs, ILDs* and *cross correlation*, as are currently provided by the Two!Ears system. Notwithstanding, WP4 concentrates on the cognitive aspects of Two!Ears and thus operates on the *symbolic level*. Herein, the formation of *auditory objects* becomes of utmost importance for scene understanding based on the ear signals of the robot. However, auditory-object formation in the sense of WP4 requires *binding* [88] of source locations and source identities. Unfortunately, this are not yet available in the current version of the Two!Ears framework but planned to be included in future versions.

To address this issue, the required auditory binding process is *emulated* in BEFT by integrating ground-truth information from the synthetic scene. The resulting auditory objects are subsequently employed for higher-level cognitive processing and action planning in moderately complex scenes. As a positive side effect, experiments run significantly faster in emulated scenarios, because auralization and auditory-feature extraction are skipped. Further, emulation supports setting up experiments easily and fast, which are not easily realized otherwise.

The following section gets granular on binding, symbol generation, and active exploration, performed in the *emulation mode* of the *Bochum Experimental Feedback Testbed* (BEFT).

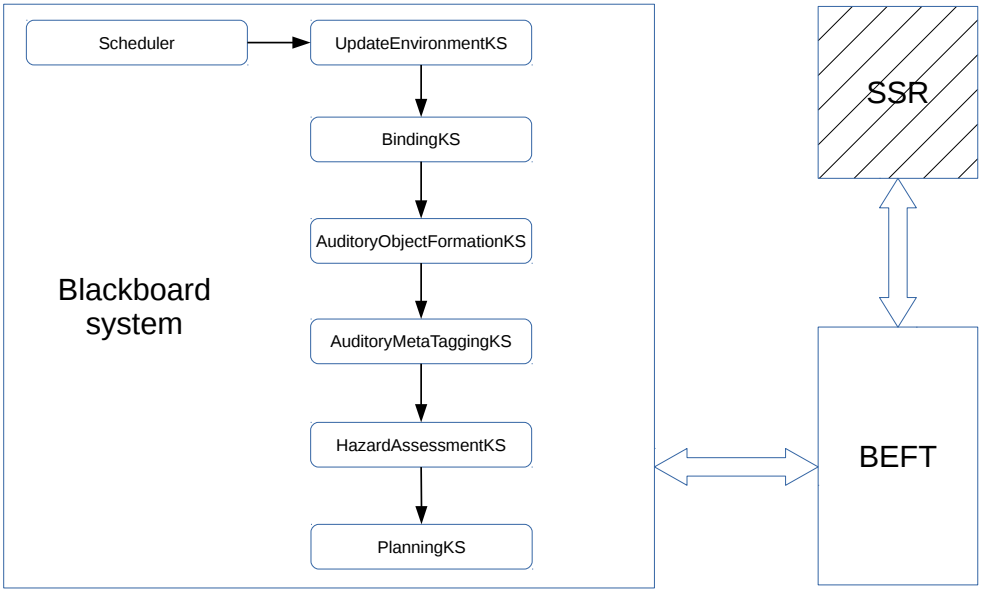


Figure 5.4: Overview of the experimental Blackboard architecture as employed in Sec. 5.4. **Blueish arrows** indicate the flow of information, and symbolize issued control commands. **Black arrows** clarify the activation sequence of the depicted knowledge sources. The **hatched depiction** of the SSR symbolizes that this system block is not used in the emulation mode of BEFT

5.2 Emulating auditory-object formation

As stated above, the "Wheres" and "Whats" of all overheard sound sources have to be combined by appropriate binding mechanisms in order to achieve successful auditory-object formation. To that end, define a *BindingKS* that receives ground-truth scenario information from the BEFT core and generates, on a per-frame basis, *auditory-object hypotheses* that correspond to all currently active sound sources.

BindingKS Let $\mathbf{p}_s = [x_s, y_s]'$ (where $s = 1...N_S$) be the x/y -plane-based ground-truth positions of all N_S sound sources in a given scenario. In addition, let l_s represent the ground-truth *label* (or *identity*) of source s. The set of all active sources in frame f be $\mathcal{E}_f$. Further, let $\mathbf{r}_f = [x_f, y_f, \phi_f]$ be the artificial robot's head pose in frame f, where x_f and y_f represent the head's position in planar world coordinates, while ϕ_f indicates the current head orientation, also in world coordinates.

Now, let $\mathbf{h}_f = [cos(\phi_f), sin(\phi_f), 0]'$ define the *heading vector* of the robot in frame f, while $\mathbf{d}_f^s = [x_s - x_f, y_s - y_f, 0]'$ be the *source direction vector* spanning from the robot's

current head position to source s. Further, let $\mathbf{n}_f^s = \mathbf{d}_f^s / \left|\left|\mathbf{d}_f^s\right|\right|$ represent the normalized source direction vector. With that, the signed azimuth location of source s in the robot's head-centric coordinate system can be defined as follows.

$$\phi_f^s = \frac{180.0}{\pi} \cdot atan2 \left(\left(\mathbf{h}_f \times \mathbf{n}_f^s\right) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \mathbf{h}_f \cdot \mathbf{n}_f^s \right) , \tag{5.2}$$

where ϕ_f^s is defined in the interval $[-180°; +180°]$. By employing $\phi_f^s = \mathrm{mod}\left(\phi_f^s, 360\right)$, the definition interval is transformed to $[0°; 360°]$. This renders the azimuth estimates compatible with output from the *DnnLocationKS*, and will facilitate migration from emulation to simulation in future system versions.

Given the location estimates for all active sources in frame f, it becomes straightforward to augment those estimates with the corresponding ground-truth labels. This yields a set of *binding hypotheses*, $\mathcal{B}_f = \left\{\phi_f^s, l_s\right\}$, with $s \in \mathcal{E}_f$. The *BindingKS* then expands this set with $[x_f, y_f, \phi_f]$ and pushes the resulting structure onto the Blackboard, thus making it available to downstream knowledge sources.

AuditoryObjectFormationKS With $\mathcal{B}_f$ being available, the robotic agent has to infer and refine $\mathbf{p}_{s,f}^e$, the x/y–plane position of each sound source, s, in each frame, f. To that end, the machine switches to *patrol mode* and roams the given scenario on a fixed path. In doing so, it uses a simplistic A* path-planning scheme (modified from [73]) that integrates basic collision avoidance. Driven by the *PlanningKS* (see Sec. 5.3), patrolling is continued until all $\mathbf{p}_{s,i}^e$ have been inferred with sufficient reliability.

During patrol, the robotic agent maintains a set, $\mathcal{O}_f^A$, of *auditory-object hypotheses,* $\mathbf{o}_{s,f}^A = \left\{\mathbf{p}_{s,f}^e, l_s\right\}$. Note that any hypothesis stored in $\mathcal{O}_f^A$ represents a self-contained, expandable set of data that models (with respect to frame f) the robot's knowledge of each acoustically observed sound source, s.

From an initially empty set, $\mathcal{O}_f^A$ grows and adapts in the course of the emulation according to the following scheme. Given that $s \in \mathcal{E}_f$, $\mathcal{O}_f^A$ is appended with $\mathbf{o}_{s,f}^A$ if and only if $\mathbf{o}_{s,f}^A \notin \mathcal{O}_f^A$. If, however, $\mathbf{o}_{s,f}^A \in \mathcal{O}_f^A$, the corresponding set entry is refined with respect to $\mathbf{p}_{s,f}^e$, while the robot moves along its prescribed path. For this purpose, assume that sound source s is currently active, and use the information in $\mathcal{B}_f$ to define the planar position of the head of the robot, $\mathbf{p}_f^r = [x_f, y_f]$. Further, let $\mathbf{d}_{s,f}^e = [\cos(\phi_f^s + \phi_f), \sin(\phi_f^s + \phi_f)]$ define the *estimated source-direction vector*. This vector is used to project a line $\mathbf{l}_s = \mathbf{p}_f^r + k \cdot \mathbf{d}_{s,f}^e$, pointing from the robot's head towards the estimated position of source s. Assume line projection to repeat in all emulation frames where $s \in \mathcal{E}_f$, up to the current frame, f.

Following [86], this allows to set up

$$\mathbf{R}_s = \sum_{i=1}^{f} \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \mathbf{d}^e_{s,i} \cdot \mathbf{d}^e_{s,i}{}^T \right) \cdot [s \in \mathcal{E}_f] \ ,$$
$$\mathbf{q}_s = \sum_{i=1}^{f} \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \mathbf{d}^e_{s,i} \cdot \mathbf{d}^e_{s,i}{}^T \right) \cdot \mathbf{p}^r_i \cdot [s \in \mathcal{E}_f] \ , \tag{5.3}$$

where $[\,\cdot\,]$ represents the *Iverson bracket*. 5.3 can eventually be referenced to retrieve a least-squares estimate of $\mathbf{p}^e_{s,f}$ [86] as follows,

$$\mathbf{p}^e_{s,f} = \mathbf{R}_s{}^{\dagger} \mathbf{q}_s \, . \tag{5.4}$$

Note that the acuity of 5.4 gradually increases with increasing emulation time, as information from more frames aggregates in (5.3) and cancels out uncertainties in azimuth estimation and in the robot's self-localization mechanism.

With the above, define the *localization instability* as follows.

$$v^s_f = \left|\left| \mathbf{p}^e_{s,f} - \mathbf{p}^e_{s,f-1} \right|\right| \ , \tag{5.5}$$

for each overheard sound source, s. Note that v^s_f is smoothed over the last $N_{smooth} = \min(f, 10)$ emulation frames. The resulting *smoothed localization instability* value is henceforth termed $\tilde{v}^s_f$.

With (5.5) define

$$\tilde{V}_f = \frac{1}{|\mathcal{E}_f|} \sum_{s \in \mathcal{E}_f} \tilde{v}^s_f \ , \tag{5.6}$$

as the *averaged global position uncertainty* used in the *PlanningKS* (see Sec. 5.3). Observe that $|\,\cdot\,|$ in 5.6 denotes the *set cardinality*.

Eventually, the *AuditoryObjectFormationKS* augments each auditory object hypothesis, $\mathbf{o}^A_{s,f}$, with the corresponding smoothed localization instability value, yielding $\mathbf{o}^A_{s,f} = \mathbf{o}^A_{s,f} \bigcup \tilde{v}^s_f$. In addition, the knowledge source stores $\tilde{V}_f$ in the Blackboard memory for further processing by downstream system blocks.

AuditoryMetaTaggingKS As already stated in Sec. 5.1, WP4 focuses on the cognitive domain and operates on the symbolic level. *Meta information* required herein is provided by the *AuditoryMetaTaggingKS*. This knowledge source augments each auditory object hypothesis $\mathbf{o}^A_{s,f}$ with additional *meta tags* that define the abstract characteristics

of the corresponding sound sources, s, in emulation frame, f. This means that the meta data used here are purely emulated! The advantage of such a procedure is that it allows for performing cognitive experiments of increased complexity, even if the required meta information is currently not yet available from lower stages of the TWO!EARS framework.

Up to now, meta tags as listed in Table 5.1 are generated by the *AuditoryMetaTaggingKS*. In this table, "NA" indicates that the corresponding meta tag may not be applicable

Meta class	**Meta subclass**	**Domain**
category	human,animal,threat,alert	0..1
role	employee, rescuer, victim, fire, siren, dog	0..1
gender	male,female,NA	0..1
stress	–	$\min\left(\max\left(\mathcal{N}\left(\mu_s, 0.025\right), 1\right), 0\right)$, NA
loudness	–	$\min\left(\max\left(\mathcal{N}\left(\mu_l, 0.025\right), 1\right), 0\right)$, NA
age	–	$\min\left(\max\left(\mathcal{N}\left(\mu_a, 2\right), 100\right), 0\right)$, NA

Table 5.1: The meta tags employed in the *Bochum Experimental Feedback Testbed* (BEFT), together with their definition range (domain)

to any $\mathbf{o}^A_{s,f}$. For instance, it would be pointless to assign stress values to an auditory object of role "fire", or supply an auditory object of role "siren" with "age" information.

Note that the BEFT contains ground-truth meta information for all instantiated sound sources. Thus, each auditory object hypothesis can principally be supplied with perfect meta knowledge of the emulated environment. This would never be possible in real-world scenarios, as the meta information can only be extracted via noisy sensors, consequently resulting in imperfect assignment of data to the corresponding $\mathbf{o}^A_{s,f}$.

In the emulation mode, to account for sensor noise, the *membership scores* for all classes defined in Table 5.1 are calculated by artificially degrading ground-truth information extracted from the BEFT. For that, let the set of available categories be $\mathcal{C} = \{c_{k=1..N_{cat.}}\}$, where $N_{cat.} = |\{human, animal, threat, alert\}|$. Further assume that each emulated source, s, is of the true category c_s. The category membership scores $C^s_f(c_{i=1..N_{cat.}})$ for sound source s are then formulated according to

$$C^s_f(c_{i=1..N_{cat.}}) = \begin{cases} \mathcal{N}\left(0.9, 0.01\right) & \text{if } c_i = c_s \\ \mathcal{N}\left(0.1, 0.05\right) & \text{otherwise}\,. \end{cases} \tag{5.7}$$

Note that defining the membership scores, $R^s_f(r_{i=1..N_{rol.}})$, for role and $G^s_f(g_{i=1..N_{gen.}})$ for gender is performed in a similar manner, that is, $N_{rol.}$ and $N_{gen.}$ represent the cardi-

nalities of the role set $\mathcal{R} = \{r_{i=1..N_{rol.}}\} = \{employee, rescuer, victim, fire, siren, dog\}$, respectively the gender set $\mathcal{G} = \{g_{i=1..N_{gen.}}\} = \{male, female, NA\}$. The *AuditoryMetaTaggingKS* then integrates all emulated membership scores into the auditory object corresponding to the specific sound source s, yielding $\mathbf{o}^A_{s,f} = \mathbf{o}^A_{s,f} \bigcup \{C^s_f, R^s_f, G^s_f\}$.

In addition to the meta information on categories, roles, and genders, let S^s_f define the stress level for sound source s in frame f. Similarly, assume loudness level and age to be defined by L^s_f, and A^s_f, respectively. These figures are then used to further expand the auditory object relating to sound source s, according to $\mathbf{o}^A_{s,f} = \mathbf{o}^A_{s,f} \bigcup \{D^s_f, L^s_f, A^s_f\}$.

HazardAssessmentKS Given the above meta information, BEFT emulates individual *hazard scores*, $H_{s,f}$, for all auditory objects in $\mathcal{O}^A_f$. Focusing on a dedicated $\mathbf{o}^A_{s,f}$, let

$$\hat{c}^s_f = \arg\max_{c_i \in \mathcal{C}} C^s_f(c_i), \quad \hat{r}^s_f = \arg\max_{r_i \in \mathcal{R}} R^s_f(r_i), \quad \hat{g}^s_f = \arg\max_{g_i \in \mathcal{G}} G^s_f(g_i) \tag{5.8}$$

be the most probable category, role, and gender of the corresponding sound source s in emulation frame, f. Further, be $d_R = \left|\left|\mathbf{d}^s_f\right|\right|$ the distance from sound source, s, to the head of the robot head.

Assuming $\mathbf{o}^A_{s,f} \in \mathcal{O}^A_f$, set up a *rescue score* according to

$$S^R_{s,f} = \max_{\mathbf{o}^A_{i,f} \in \mathcal{O}^A_f} \left[R^i_f(\text{“rescuer”}) \cdot \exp\left(-\frac{\left|\left|\mathbf{p}^e_{s,f} - \mathbf{p}^e_{i,f}\right|\right|^2}{4} \right) \right] . \tag{5.9}$$

High values of 5.9 may induce, for example, one of the following two hypotheses. Either, the auditory object that relates to sound source s is a rescuer itself, and can thus be assumed to require only minor help from the robotic agent. Or $\mathbf{o}^A_{s,f}$ is spatially close to another auditory object that likely fulfills the “rescuer” role. In the latter case, the system expects the nearby rescuer to look after the scenario entity corresponding to $\mathbf{o}^A_{s,f}$, allowing the robot to focus its attentional resources on entities with lower rescue scores. Note that in both cases, higher $S^R_{s,f}$ cause $H_{s,f}$ to lower – see (5.11).

Antagonizing (5.9), set up a *threat score*, $S^T_{s,f}$, for the auditory objects in $\mathcal{O}^A_f$, such that

$$S^T_{s,f} = \max_{\mathbf{o}^A_{i,f} \in \mathcal{O}^A_f} \left[C^i_f(\text{“threat”}) \cdot \exp\left(-\frac{\left|\left|\mathbf{p}^e_{s,f} - \mathbf{p}^e_{i,f}\right|\right|^2}{4} \right) \right] , \tag{5.10}$$

The term (5.10) takes on high values if the entity represented by $\mathbf{o}^A_{s,f}$ is expected to be spatially close to another auditory object that likely belongs to the "threat" category. In such a dangerous situation, the attention of the robot has to focus on the threatened entity, resulting in an increased $H_{s,f}$ – see 5.11. With the above, the individual hazard score is assembled according to

$$H_{s,f} = S^s_f + L^s_f + R^s_f(\text{“victim”}) - S^R_{s,f} + S^T_{s,f}\,. \tag{5.11}$$

The term 5.11 enforces increasing values for $H_{s,f}$ if the scenario entity corresponding to the auditory object $\mathbf{o}^A_{s,f}$

- Is likely to be a victim
- Shows increased voice stress S^s_f or loudness L^s_f
- Is far from a rescuer
- Is close to a threat (e.g. fire)

During rescue attempts, all animate beings have to be evacuated from the scenario. However, as humans have to be rescued first, the individual threat score is post-processed using

$$H_{s,f} = 0.25 \cdot H_{s,f} \qquad \text{if } \hat{c}^s_f = \text{“animal”}\,. \tag{5.12}$$

In the experiments descibed below, it is further assumed that inanimate entities cannot be threatened, requiring a further post-processing step according to

$$H_{s,f} = 0 \qquad \text{if } \hat{c}^s_f \in \{\text{“threat”}, \text{“alert”}\}\,. \tag{5.13}$$

While quite basic, the above ad-hoc definition of the individual hazard scores results in reasonable behavior of the robotic agent in the DASA experiments discussed below. Notwithstanding that, (5.11) can readily be extended to suit the demands of more complex, future DASA scenarios. For upcoming versions of Two!Ears, it would also be possible to have human assessors judge on hazard scores in a variety of emulated DASA situations. Data recorded from these trials could then be used to train a deep neural network that infers $H_{s,f}$ directly and replaces the ad hoc solution proposed above.

To conclude discussion of the *HazardAssessmentKS*, note that the individual hazard scores are averaged over a sliding time window of the last 30 frames, resulting in the *smoothed individual-hazard scores,* $\tilde{H}_{s,f}$. With E^+_f being the number of entities for which $\tilde{H}_{s,f} > 0$, the *global hazard score* is calculated according to

$$H^G_f = \frac{1}{E^+_f} \sum_{\mathbf{o}^A_{i,f} \in \mathcal{O}^A_f} \tilde{H}_{i,f}\,. \tag{5.14}$$

The time course of the global hazard score is memorized up to the current frame, f. This results in an array $\mathbf{T}_H = \left[H_1^G, ...H_f^G\right]$. With that define

$$M_f^G = \frac{1}{|\mathbf{T}_H|} \sum_{i=1}^{|\mathbf{T}_H|} H_i^G \ . \tag{5.15}$$

The value provided by 5.15 is then passed on to the Blackboard for use in the *PlanningKS*descibed below.

5.3 A knowledge source for action planning

The knowledge sources defined above act together in order to allow for *bottom up* scenario analysis. However, without *top-down* active exploration, the *AuditoryObjectFormationKS* would not be able to localize overheard sound sources. In turn, computation of the hazard scores as provided by the *HazardAssessmentKS* would be flawed. Consequently, it would be impossible to derive a meaningful action plan for the robotic rescuer without adequate scenario understanding driven by top-down mechanisms.

To account for the insights mentioned above, the Blackboard architecture is augmented with a cognitive expert subsystem, the *PlanningKS*. This knowledge source enables meaning assignment and scene understanding together with high-level planning as well as active scenario exploration. The *PlanningKS* can be seen as the "brain" of the robotic agent. Realized as a task stack, The *PlanningKS* employs a manually derived rule set to issue new tasks and provide meaningful robotic behavior in moderately complex scenarios.

To be sure, the internal architecture of the *PlanningKS* has to be adaptable to novel situations. For instance, the understanding and handling of the DASA situation discussed in (5.4) is enabled through the rules and tasks encoded in Fig. 5.5. Only the major tasks have been depicted in this figure, minor subtasks like actuator control have been left out for clarity.

In future system versions, manual adaptation of the *PlanningKS* could be automatized as follows. On the one hand, deep neural network methods could be used to infer purposeful robotic action plans directly from data collected in human trials. On the other hand, application of reinforcement-learning techniques [83] could be employed to enable the robot to discover reasonable action patterns in a completely autonomous manner.

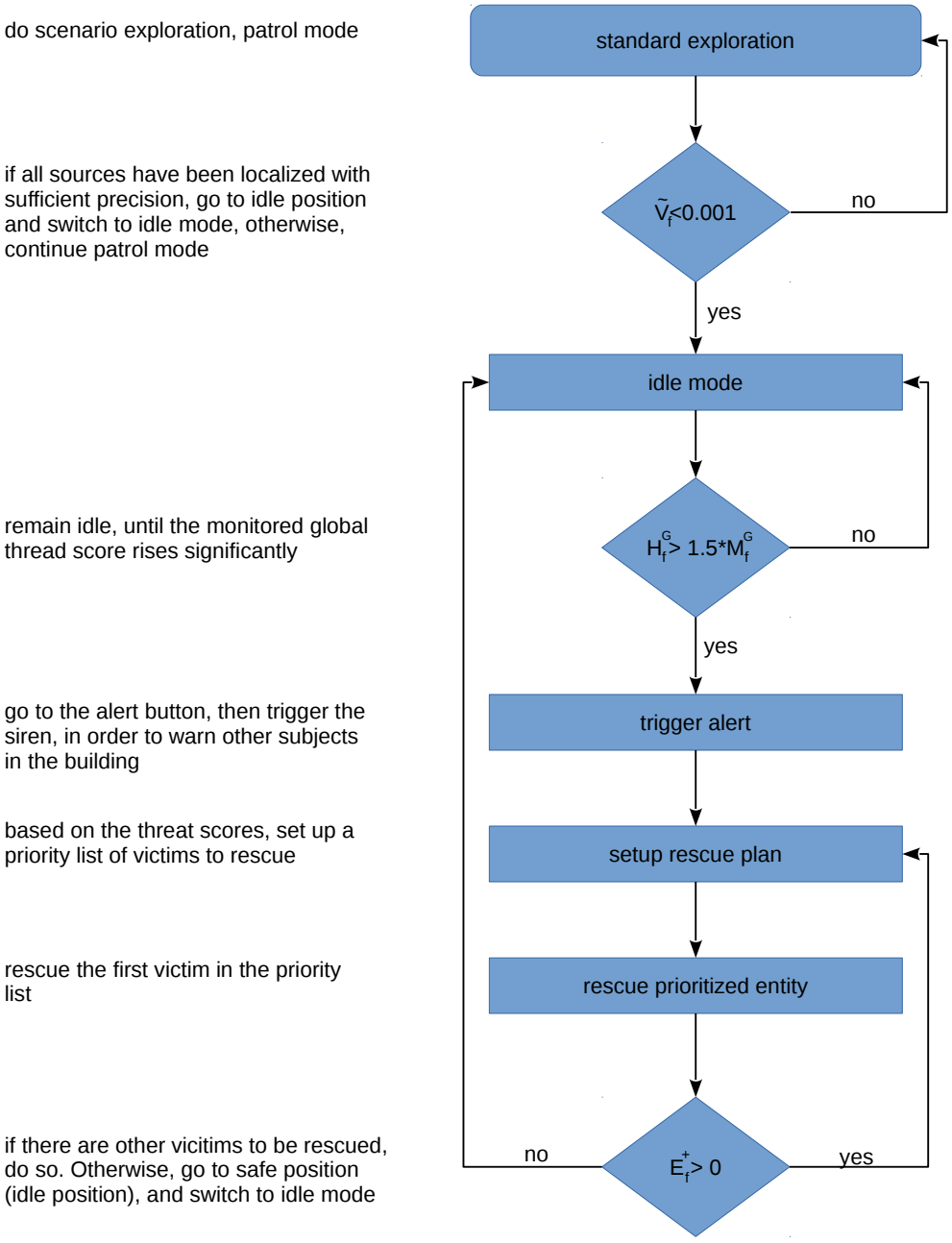


Figure 5.5: Overview of the *PlanningKS* architecture employed in the cognitive experiment described in Sec. 5.4. The flow plan on the right subsumes the tasks **(rectangles)** and decision rules **(diamonds)** embodied in the framework for scene understanding and active exploration. The hints on the left provide details of the corresponding diagram blocks

5.4 Instrumental evaluation

To qualitatively assess the capabilities of BEFT in an emulated dynamic auditory scene, a simplified version of the ADREAM lab – which is the "computer appartment" set up at LAAS in Toulouse – has been replicated using the 3–D-modeling capabilities of Blender [12]. The emulation of the scenario in BEFT has a duration of 400 s. It starts in normal lab conditions, then, after $T_{event} = 60$ seconds, the situation evolves into a catastrophic scenario. After an assumed explosion, attendant lab employees turn into victims or rescuers, and a fire starts in one corner of the lab. Table 5.2 subsumes the meta characteristics of all entities present in the proposed scenario, including their roles before and after T_{event}.

Entity	**category**	**pre event role**	**post event role**	**gender**	**age**
Source001	human	employee	victim	male	25
Source002	animal	dog	victim	male	2
Source003	human	employee	rescuer	female	30
Source004	human	employee	victim	male	40
Source005	alert	siren	siren	NA	NA
Source006	threat	fire	fire	NA	NA
Source007	human	employee	victim	female	20

Table 5.2: Meta characteristics of the entities present in the cognitive experiment discussed in Sec. 5.4

Each of the entities in the table above corresponds to an emulated sound source, s. Let U_s be an *utterance schedule* that contains the emission pattern and potential role changes for each entity. Note that the emulation mode of BEFT does not require the availability of physical stimuli connected to the utterances stored in each U_s. This allows to define entities of nearly arbitrary category and role without the need for huge sound databases. Fig. 5.6 shows the activation pattern for each sound source in the current context. Emission and silence intervals are guided by noise distributions with programmed means and standard deviations.

Prior to T_{event}, all animate entities display a low stress level, indicated by their vocal activities. The emulated sound sources corresponding to the "fire" and "siren" entities remain inactive. After T_{event}, the stress level suddenly rises in the animate entities. "Source001" is assumed to become unconscious after several seconds, ceasing its emission. "Source004" is expected to be severely injured, causing the loudness level of its utterances to drop significantly. Note also that although the emission of the siren is theoretically available at T_{event}, the activation of the corresponding entity is deliberately postponed until the "trigger alert" step in Fig. 5.5 is executed.

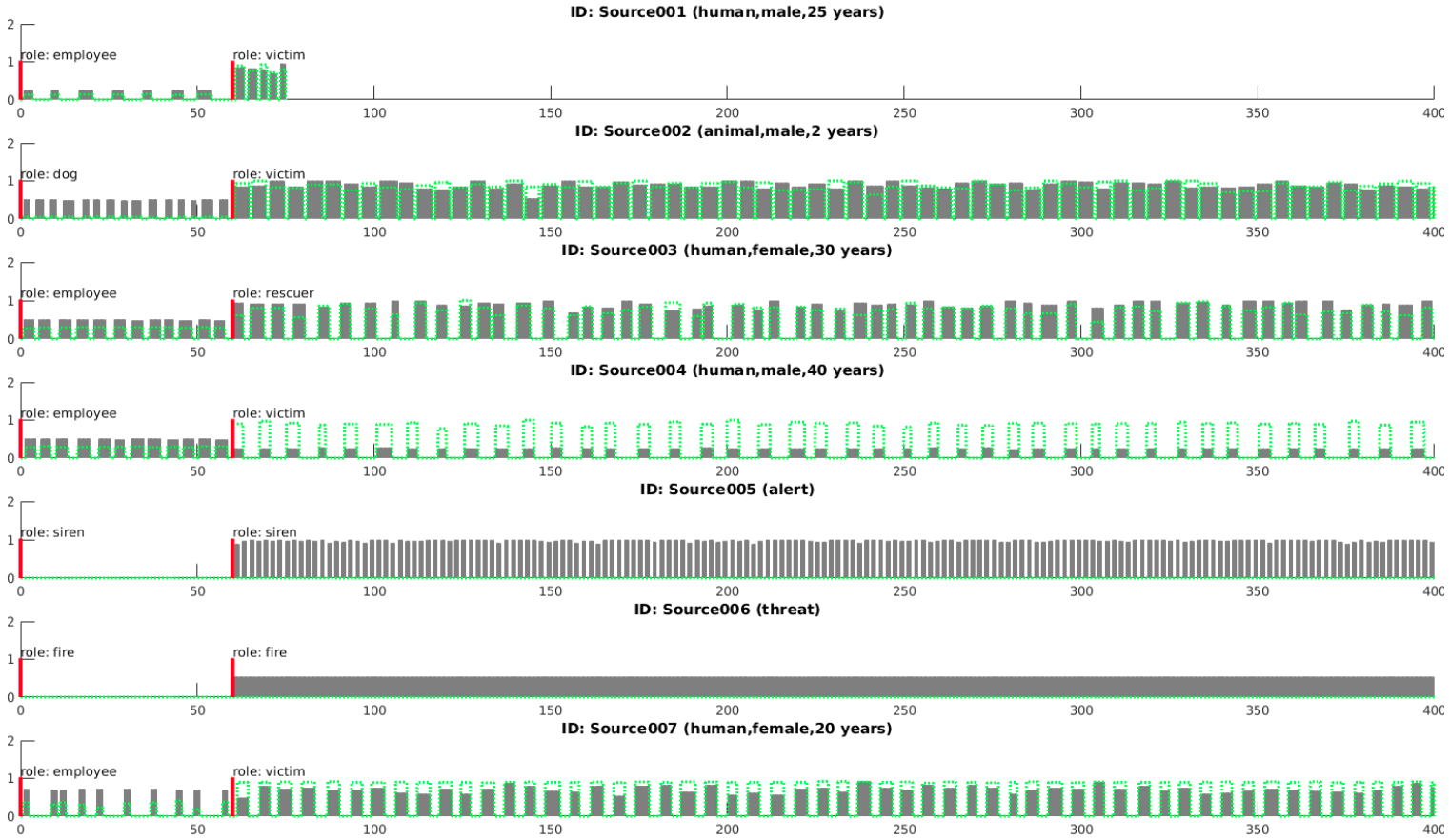


Figure 5.6: Overview of the source activation schedule employed in the cognitive experiment described in Sec. 5.4. **Grey bars** indicate the emulated loudness of each stimulus, the **green dotted line** traces the emulated stress level of each entity. The *leftmost red bar* indicates the start of the emulation, the following **red bar** marks T_{event}

Figure 5.7: The experiment described in Sec. 5.4 seen from BEFT's point-of view. **Yellowish boxes** indicate active sound sources, the **blueish rectangle** is the idle position of the robot. Identities have been annotated manually for illustration. In the depicted state of the emulation, the robotic agent rescues the prioritized entity with label "Source007"

Given the information encoded in Fig. 5.6, the robotic agent uses the logic circuitry defined in Fig. 5.5 to sense the upcoming catastrophic conditions at T_{event} and, eventually, to evacuate all animate entities from the scene. Figure 5.7 shows the BEFT's view of the created scene. The avatars were either hand-crafted or designed using *MakeHuman*^{TM} [66], a 3–D-modeling tool for human characters. Wall proxies in the scene were derived from project-internal measurements done in the ADREAM lab. Note that in the current version of BEFT, the walls acousticly act as "ghosts", that means that no wall reflections are emulated. For methods of how to include the acoustic and psychoacoustic effects, refer to the *Precedence effect* – compare Chap. 4, Sec. 4.2.1.

To visualize the internal model of the world that the robot has developed for itself, Fig. 5.8 shows the *environmental map* containing the estimated locations of all entities (green circles). Filled circles represent reliable source estimates, empty circles symbolize sources

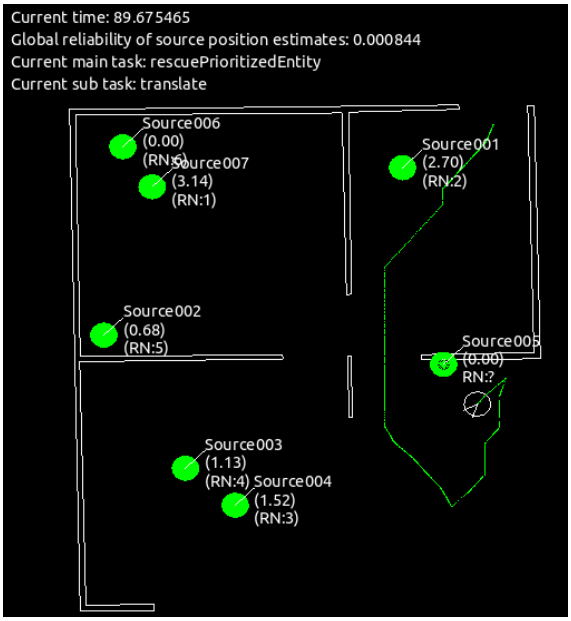


Figure 5.8: The environmental map that represents the robot's internal model of the environment

with high localization instability. The smoothed individual hazard scores and the rescue-sequence numbers (RNs) for each sound source are annotated next to the position estimates. Lower RNs indicate higher priority for being rescued. Note that inanimate entities display zero hazard scores and will not be addressed during evacuation. In addition to these data, the emulation time line and the global hazard score are shown at the top, together with the task/sub task currently performed by the robot.

BEFT allows to automatically generate a range of different scenarios with varying characteristics, thus allowing for quantitative assessment of the performance of *search and rescue* (S-&-R) schemes encoded in the *PlanningKS*. Focusing on the S-&-R strategy discussed in Fig. 5.5, $N_R = 30$ scenarios like the one depicted in Fig. 5.9 are generated by randomly altering the x–y-positions of all animate entities. Green cylinders in Fig. 5.9 represent all $\mathbf{p}_s$, with $s = 1...N_S$. The red and yellow patches symbolize "forbidden" areas where no animate entity can be placed. In this way, the randomly positioned sources are kept away from walls and become unlikely to completely stall the robot during scenario exploration.

Let the current emulation time, T_r^A, correspond to the moment where $\tilde{V}_f$ drops below 0.001 in scenario r – see Fig. 5.5. In addition, be T_r^B the time span required to evacuate all animate entities from scenario r. This allows to define the arithmetic

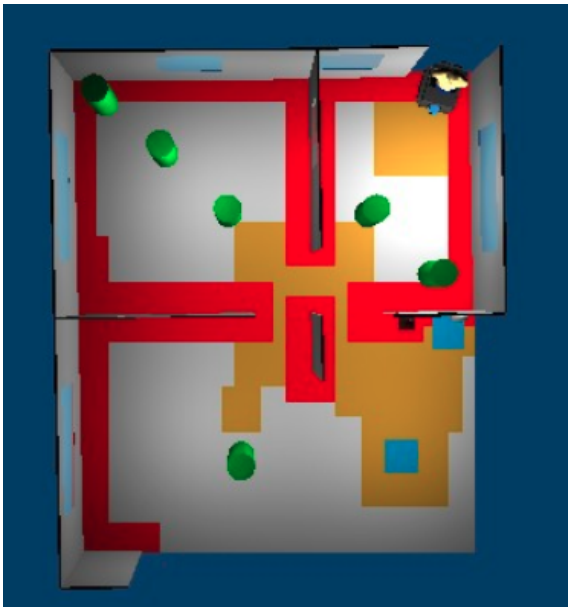


Figure 5.9: Source placement in a sample S-&-R scenario

means

$$\mu_A = \frac{1}{N_R}\sum_{r=1}^{N_R} T_r^A, \qquad \mu_B = \frac{1}{N_R}\sum_{r=1}^{N_R} T_r^B \tag{5.16}$$

and the corresponding standard deviations as

$$\sigma_A = \sqrt{\frac{1}{N_R}\sum_{r=1}^{N_R} \left(T_r^A - \mu_A\right)^2}, \qquad \sigma_B = \sqrt{\frac{1}{N_R}\sum_{r=1}^{N_R} \left(T_r^B - \mu_B\right)^2}. \tag{5.17}$$

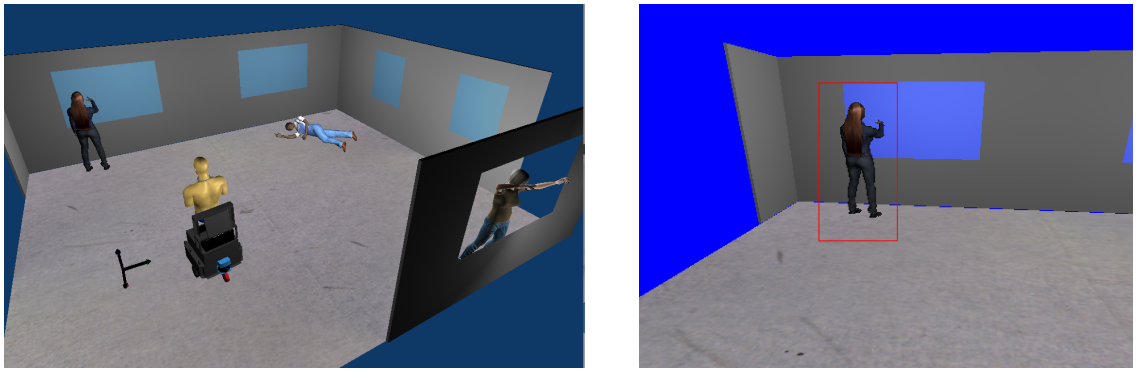
In the current experiment the obtained values are $\mu_A = 36.8609\,\mathrm{s}$, $\sigma_A = 6.1922\,\mathrm{s}$, and $\mu_B = 248.6996\,\mathrm{s}$, $\sigma_B = 31.6340\,\mathrm{s}$. In upcoming experiments, these values will have to be compared with results from trials where human assessors guide the robotic agent manually through numerous emulated rescue attempts. This would also set the pace for perceptual evaluation in addition to the instrumental one applied so far.

5.5 Multi-modal cue integration

The Two!Ears framework is aimed at multi-modal augmentation of auditory scene understanding. To this end, the physical robotic agent is equipped with a binocular camera

system that enables the capturing of video footage in given scenarios. Visual cues extracted from the image streams that the cameras deliver can be used to complement incoming auditory information, thereby enhancing the robot's comprehension of the explored environment.

To assess the benefits of audio-visual cue integration within the emulation framework, the virtual robotic agent in BEFT is equipped with a monocular camera that allows to capture the robot's field of view. Note that artificial stereo vision is not used in the current experiments, yet can readily be integrated in upcoming system versions. Henceforth, assume $\mathbf{I}_f$ to represent a virtual camera image captured in emulation frame f. Figure 5.10a shows a birds-view representation of an emulated scenario. Figure 5.10b depicts the perspective, $(\mathbf{I}_f)$, that the robot has of the same scene.



(a) **(b)**

Figure 5.10: Synthetic vision in the *Bochum Experimental Feedback Testbed* (BEFT). **(a)** A baseline scenario for assessing the effects of multi-modal-cue integration. **(b)** The same scenario as seen from the perspective of the robot. The **red box** indicates detection of an upright person

The scenario sketched in Fig. 5.10a modifies the lab model employed in Sec. 5.4 and constitutes the current basis for assessing audio-visual cue integration in BEFT. Three victims have to be rescued from the emulated scene, where an assumed $T_{event} = 0$ forces the robot into rescue mode ab initio.

The entities enrolled in the given scenario correspond to a subset $\{s_1 =$ "Source003", $s_2 =$ "Source004", $s_3 =$ "Source007"$\}$ of the sources defined in Table 5.2, with their individual roles switched to "victim". All entities are in panic, causing nearly identical stress levels, $S_f^1 \approx S_f^2 \approx S_f^3$. The procumbent male victim, s_2, is supposed to be severely injured, with his utterances significantly muffled. In contrast, the entities related to s_1 and s_3 display physical integrity and are assumed to actively yell for help. Thus one gets $L_f^2 << L_f^1$ and $L_f^2 << L_f^3$ for all emulation frames, f.

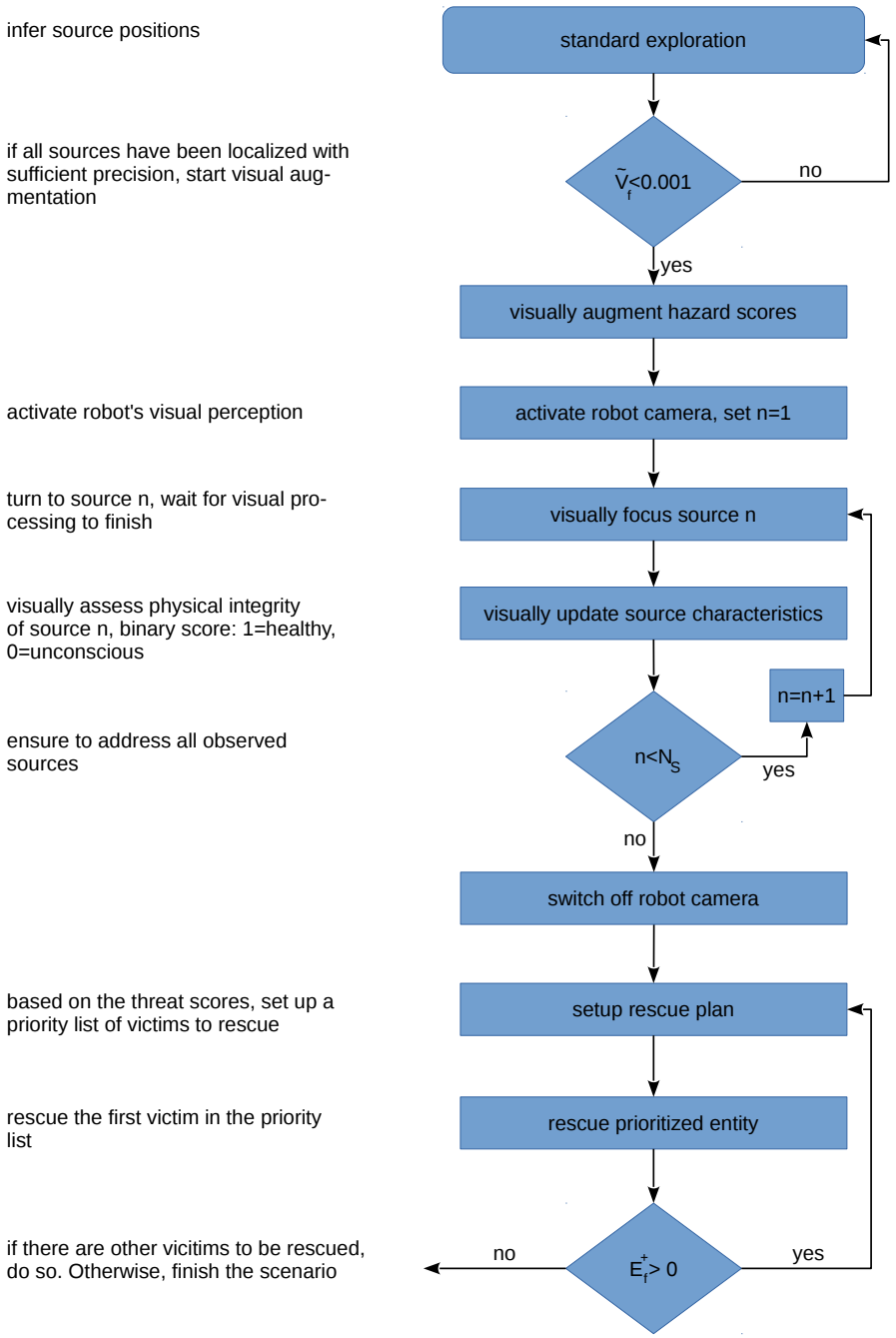
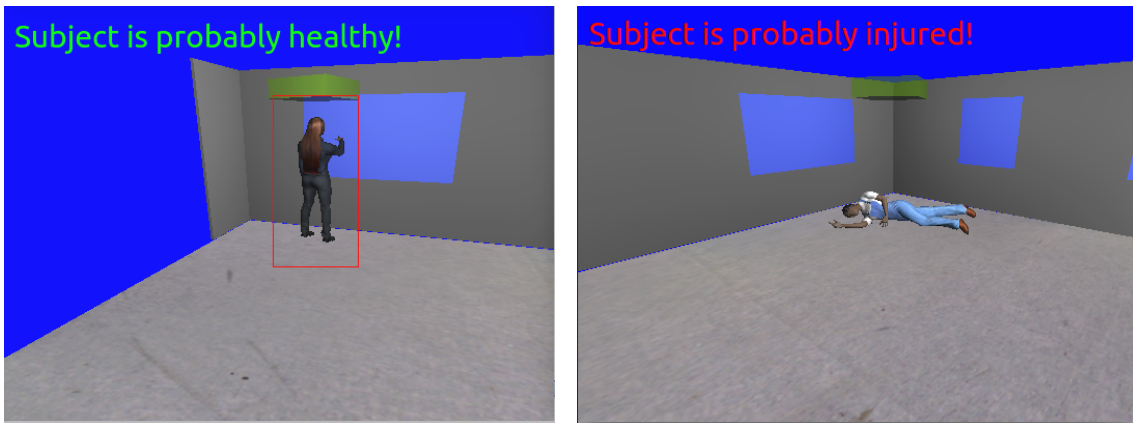


Figure 5.11: Overview of the *PlanningKS* architecture employed in the cognitive experiment described in Sec. 5.5. The flow plan on the right subsumes the tasks **rectangles** and decision rules **diamonds** embodied in the framework for audio-visual scene understanding and active exploration. The hints on the left provide details of the corresponding diagram blocks



(a) **(b)**

Figure 5.12: Visual assessment of the physical integrity of emulated victims. **(a)** A victim has been focused, the HOG detector reports presence of an upright person **(red box)**. From that, it can be deduced that the victim is fully conscious and physically integer. **(b)** Another victim has been focused, yet the HOG detector shows a negative response. Thus, the observed entity is likely to be injured and probably dizzy. The image annotations are automatically generated by BEFT. **Yellowish blocks** indicate acoustic source activity

Since there are no rescuers or threats present in the given scene, the hazard scores computed for the entities become directly proportional to their individual loudness levels – see (5.11). As a consequence, the robot would at first evacuate the active, intact entities corresponding to sound sources s_1 and s_3, disregarding the helpless person represented by s_2. Such behavior counters human intuition and is clearly inadmissible.

To suppress this inadequate behaviour, visual information from incoming $\mathbf{I}_f$ can be exploited in the following way. A *Histogram of Oriented Gradients* (HOG) detector [27] from the *OpenCV* [93] library is used to detect upright persons in $\mathbf{I}_f$. With this additional information, the robot's rescuing pattern as defined by the *PlanningKS* is adapted – see Fig. 5.11. Once the robot acoustically locked the positions of all sound sources with sufficient reliability, it activates its camera and focuses sequentially the estimated individual source locations, $\mathbf{p}_s$ for all $s = 1..N_S$. The vision-based *physical integrity* of the victim , P^s_{vis}, corresponding to sound source s, can be assessed – see Fig. 5.12.

As a result, if the robot's heading is geared towards $\mathbf{p}_s$, and the HOG detector reports the visual presence of an upright person (Fig. 5.12a), it is assumed that the focused victim is fully conscious and physically integer, thus yielding $P^s_{vis} = 1$. On the contrary, if the HOG detector displays a negative response (see Fig. 5.12b), the victim at $\mathbf{p}_s$ is deemed severely injured, and probably dizzy, causing $P^s_{vis} = 0$. By the way, the computational

load induced in the HOG-detection scheme is kept at bay by shutting down the camera of the robot unless visual augmentation is demanded.

BEFT memorizes the physical-integrity values for all assessed entities for using them to update the characteristics of the corresponding sound sources, according to

$$H_{s,f} = 5.0 \cdot H_{s,f} \qquad \text{if } P^s_{vis} < 1.0\,. \tag{5.18}$$

By (5.18), the hazard score for all unconscious victims is significantly boosted, causing helpless entities to become prioritized in the build-up of a rescue plan for the given scenario. Then the procumbent person corresponding to "Source004" will be evacuated first, resulting in a behavior of the robotic agent that matches human intuition.

5.6 Effects of illumination variations

In Sec. 5.5, a perfectly homogeneous illumination has been assumed in the acquisition of auxiliary visual information. However, this assumption falls short of real-world scenarios, as environmental lighting is generally much more complex there, which can significantly impact visual-cue extraction.

ReliabilityOfVisualCuesKS Accounting for the above insight, the Blackboard architecture shown in Fig. 5.13 integrates the *ReliabilityOfVisualCuesKS* . This knowledge source exploits image intensities to estimate the *reliability* , R^{vis}_f, of visual cues extracted from incoming camera footage, $\mathbf{I}_f$. To this end, let $\mathbf{I}^G_f$ be an *intensity image* resulting from a gray-level conversion of $\mathbf{I}_f$. The *averaged, normalized intensity* of $\mathbf{I}^G_f$ is defined by

$$B_f = \frac{1}{255.0} \cdot \frac{\sum\limits_{i=1}^{|\mathbf{I}^G_f|} \mathbf{I}^G_{i,f}}{\left|\mathbf{I}^G_f\right|}\,, \tag{5.19}$$

where $\left|\mathbf{I}^G_f\right|$ is the total number of pixels in $\mathbf{I}^G_f$, while $\mathbf{I}^G_{i,f}$ represents the i^{th} pixel in the intensity image. Note that $B_f \to 0$ (image too dark), as well as $B_f \to 1$ (image overexposed) indicate that $\mathbf{I}^G_f$ becomes useless for visual-cue extraction. As a consequence, R^{vis}_f is defined by the plateau function -0.25cm

$$R^{vis}_f = \frac{1}{1 + e^{\left(-\beta\left(B_f - \alpha_1\right)\right)}} \cdot \frac{1}{1 + e^{\left(\beta\left(B_f - \alpha_2\right)\right)}} \tag{5.20}$$

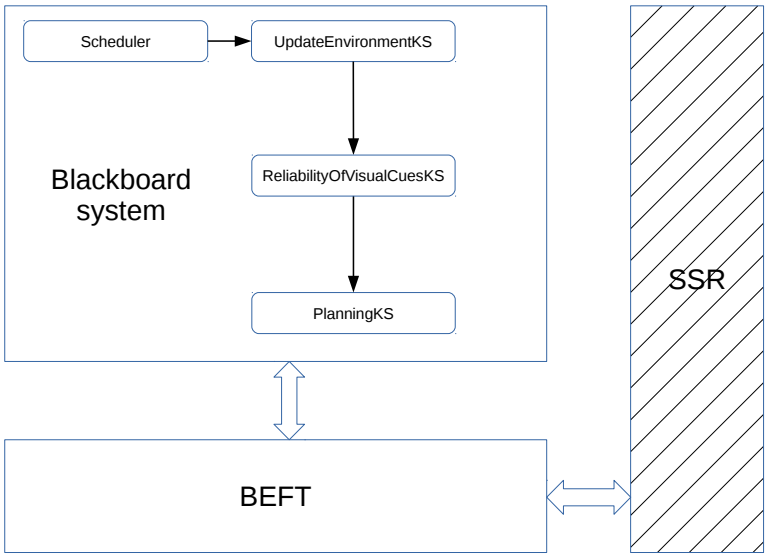


Figure 5.13: Overview of the Blackboard architecture employed in the cognitive experiment in Sec. 5.6. **Blueish arrows** indicate the flow of information and symbolize issued control commands. **Black arrows** clarify the activation sequence of the depicted knowledge sources. The **hatched depiction** of the SSR in the setup means that this system block is not used by BEFT in emulation mode

where $\beta = 100$, $\alpha_1 = 0.15$, and $\alpha_2 = 0.85$ were chosen empirically. (5.20) is depicted in Fig. 5.14. As postulated, values close to the extrema of B_f cause R_f^{vis} to drop to zero, thereby indicating potential issues in visual-cue extraction.

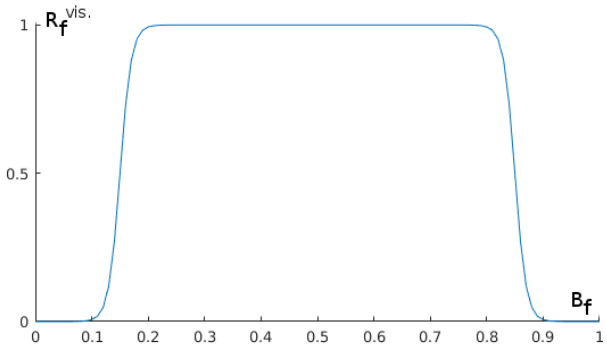
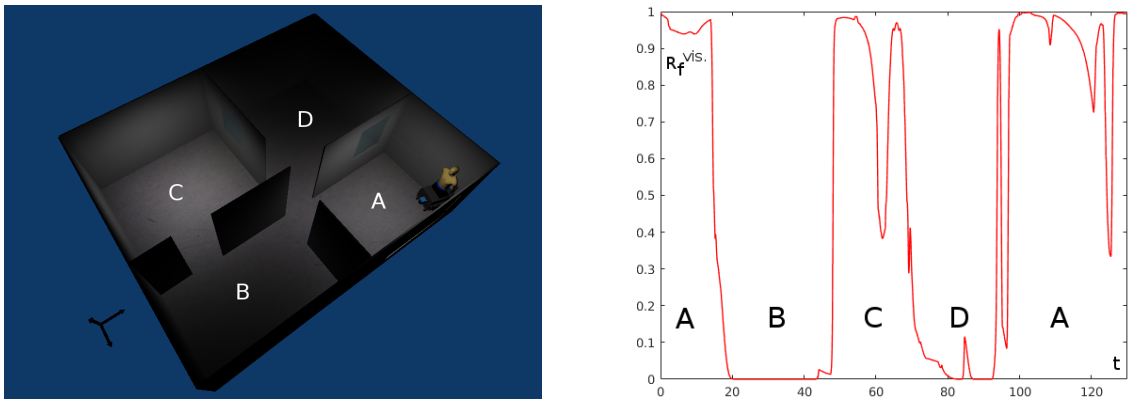


Figure 5.14: Graphical display of the visual-cue reliability defined in 5.20. Note the "forbidden" zones close to the extrema of B_f. In these zones images are either too dark or too light for visual-cue extraction

Instrumental evaluation To instrumentally assess the effects of illumination variations on R_f^{vis}, the baseline scenario shown in Fig. 5.10 is set up. No sound sources are present in this environment. The infrastructure is derived from the lab model employed in Sec. 5.4. Note that illumination conditions are strongly heterogeneous, that is, lights in rooms A and C are switched on, while the rooms B and D remain unlighted. The robot explores the terrain following a predefined unlighted path, namely, {room A → room B → room C → room D → room A}. The time course of R_f^{vis} is logged for all emulation frames, yielding the plot in Fig. 5.15b. In the well-lit rooms, A and C, R_f^{vis} (red curve) approaches values close to one, hypothesizing that visual cues extracted in these compartments are reliable. On the contrary, the gloomy rooms, B and D, cause the visual reliability measure to approach zero, indicating that the provided illumination is insufficient for reliable visual-cue extraction. In conclusion, it can be stated that R_f^{vis} behaves as expected in the proof-of-concept scenario, as shown in Fig. 5.10. Future system versions could enhance the current, basic definition of the visual-reliability measure (5.20) and eventually employ it in complex scenarios to inhibit the extraction and processing of flawed visual cues.

Softwarestatus

Data/algorithm available: yes
Code written and tested: yes
Implemented on TWO!EARS: A light version (LVTE) is been integrated. The full BEFT version is also available, but quite complex. Potential



(a) **(b)**

Figure 5.15: Reliability of visual cues. **(a)** The robot traverses a scenario without sound sources. Illumination conditions in the separate rooms vary significantly, lights in rooms A and C are switched on, room B and D are not illuminated. **(b)** R_f^{vis} over time **(red curve9**. Letters correspond to the room labeling in Fig. 5.15a. Peaks of R_f^{vis} are observed when the robot explores rooms A and C

users are advised to contact Thomas Walther (RUB)
Runs on the robot: not intended

Bibliography

[1] Aaronson, N. and Hartmann, W. (**2014**), "Testing, correcting, and extending the Woodworth model for interaural time difference," *The Journal of the Acoustical Society of America* **135**, pp. 817–823. (Cited on pages 53 and 57)

[2] Allen, J. B. and Berkley, D. A. (**1979**), "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.* **65**, pp. 943–950. (Cited on page 30)

[3] Argentieri, S., Cohen-L'Hyver, B., Danès, P., Dollat, X., Forgue, T., Gas, B., Herrb, M., Mallet, A., Manhès, J., Musabini, A., Piat, J., Podlubne, A., and Vandeportaele, B. (**2016**), *D5.3: Final Report on Hardware/Software Integration and Robotics Test Bed*, chap. 5. (Cited on page 61)

[4] Arnal, L. H. and Giraud, A.-L. (**2012**), "Cortical oscillations and sensory predictions." *Trends in cognitive sciences* **16**(7), pp. 390–8, URL `http://www.ncbi.nlm.nih.gov/pubmed/22682813`. (Cited on page 72)

[5] Baranes, A. and Oudeyer, P.-Y. (**2010**), "Intrinsically Motivated Goal Exploration for Active Motor Learning in Robots: A Case Study," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pp. 1766–1773. (Cited on page 72)

[6] Barker, J., Cooke, M., and Ellis, D. (**2002**), "Temporal integration as a consequence of multi-source decoding," in *ISCA Workshop on the Temporal Integration in the Perception of Speech*. (Cited on page 14)

[7] Benjamin, C.-l., Sylvain, A., and Bruno, G. (**2016**), "Multimodal fusion and inference using binaural audition and vision Multimodal fusion and inference using binaural audition and vision," in *International Congress on Acoustics*. (Cited on page 63)

[8] Bishop, C. (**2006**), *Pattern Recognition and Machine Learning*, Springer. (Cited on page 89)

[9] Blauert, J. (**1997**), *Spatial Hearing – The psychophysics of human sound localization, 2nd edition*, MIT-Press, (expanded and revised edition of: RÃďumliches HÃűren, S. Hirzel Verlag, DâĂŞStuttgart, 1974). (Cited on page 91)

[10] Blauert, J. and Cobben, W. (**1978**), "Some Consideration of Binaural Cross Correlation Analysis," *Acustica* **39**, pp. 96–104. (Cited on page 17)

[11] Blauert, J. and Obermayer, K. (**2012**), "Rückkopplungswege in Binauralmodellen Feedback paths in binaural models," in *Fortschritte der Akustik, DAGA '2012*, Dtsch. Ges. Akustik, D-Berlin, pp. 2015–2016. (Cited on page 5)

[12] Blender Foundation (**2014**), "Blender - 3D open source animation suite," URL `http://www.blender.org/`. (Cited on pages 8, 87, and 100)

[13] Braasch, J. (**2013**), "A precedence effect model to simulate localization dominance using an adaptive, stimulus parameter-based inhibition process," *J. Acoust. Soc. Am.* **134**(1), pp. 420–435. (Cited on pages 18, 24, 32, and 40)

[14] Braasch, J. (**2016**), "Binaurally Integrated Cross-correlation Auto-correlation Mechanism (BICAM)," *J. Acoust. Soc. Am. (Express Letter)* **140**(1), pp. EL143âĂŞ–EL148. (Cited on pages 26, 28, and 30)

[15] Braasch, J. and Blauert, J. (**2003**), "The precedence effect for noise bursts of different bandwidths. II. Comparison of model algorithms," *Acoust. Sci. & Tech.* **24**, pp. 293–303. (Cited on pages 17 and 19)

[16] Braasch, J., Clapp, S., Parks, A., Pastore, T., and Xiang, N. (**2013**), "A binaural model that analyses aural spaces and stereophonic reproduction systems by utilizing head movements," in *The Technology of Binaural Listening Application of Models of Binaural Listening Binaural Listening in Technology*, edited by J. Blauert, Springer, Berlin, Heidelberg, New York, pp. 201–223. (Cited on pages 26, 27, and 31)

[17] Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**), "Binaural processing model based on contralateral inhibition. I. Model setup," *J. Acoust. Soc. Am.* **110**, pp. 1074–1088. (Cited on page 34)

[18] Bronkhorst, A. W. (**2000**), "The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions," *Acta Acustica united with Acustica* **86**(1), pp. 117–128. (Cited on page 39)

[19] Brown, G. and Cooke, M. (**1994**), "Computational auditory scene analysis," *Comput. Speech. Lang.* **8**, pp. 297–336. (Cited on page 14)

[20] Bullet (**2014**), "The Bullet Physics Engine," URL `http://bulletphysics.org/wordpress/`. (Cited on page 87)

[21] Bustamante, G., Portello, A., and Danès, P. (**2015**), "A Three-Stage Framework to Active Source Localization from a Binaural Head," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'2015)*, Brisbane, Australia.

(Cited on page 52)

[22] Capdepuy, P., Polani, D., and Nehaniv, C. L. (**2007**), "Maximization of Potential Information Flow as a Universal Utility for Collective Behaviour," in *IEEE Symposium on Artificial Life*, Ieee, pp. 207–213, URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4218888`. (Cited on page 72)

[23] Cohen-Lhyver, B., Argentieri, S., and Gas, B. (**2015**), "Modulating the Auditory Turn-to Reflex on the Basis of Multimodal Feedback Loops : the Dynamic Weighting Model," in *IEEE Robio*. (Cited on page 63)

[24] Cooke, M. (**1993**), *Modelling auditory processing and organisation*, Cambridge University Press, Cambridge, MA. (Cited on page 14)

[25] Cooke, M., Barker, J., Cunningham, S., and Shao, X. (**2006**), "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.* **120**, pp. 2421–2424. (Cited on page 14)

[26] Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**), "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication* **34**(3), pp. 267 – 285, URL `http://www.sciencedirect.com/science/article/pii/S0167639300000340`. (Cited on page 47)

[27] Dalal, N. and Triggs, B. (**2005**), "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893. (Cited on page 107)

[28] David, H. A. and Nagaraja, H. N. (**2003**), *Order Statistics*, Wiley, 3 ed. (Cited on page 47)

[29] de Cheveigné, A. and Kawahara, H. (**2002**), "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.* **111**, pp. 1917–1930. (Cited on page 17)

[30] Doclo, S., den Bogaert, T. V., Moonenb, M., and Wouters, J. (**2006**), "Binaural noise reduction in hearing aids," URL `https://www.uni-oldenburg.de/fileadmin/user_upload/mediphysik/ag/sigproc/download/presentations/doclo/pres_ASIP-NET_261006.pdf`. (Cited on page 44)

[31] Duangudom, V. and Anderson, D. V. (**2007**), "Using auditory saliency to understand complex auditory scenes," in *15th European Signal Processing Conference*. (Cited on page 71)

[32] Durlach, N. I. (**1960**), "Note on the Equalization and Cancellation theory of binaural masking level differences," *J. Acoust. Soc. Am.* **32**, pp. 1075–1076. (Cited on page 34)

[33] Faller, C. and Merimaa, J. (**2004**), "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.* **116**, pp. 3075–3089. (Cited on pages 34 and 35)

[34] Friedman, J., Hastie, T., and Tibshirani, R. (**2010**), "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software* **33**(1), pp. 1. (Cited on page 47)

[35] Gaese, B. H. and Wagner, H. (**2002**), "Precognitive and cognitive elements in sound localization," *Zoology* **105**, pp. 329–339. (Cited on page 11)

[36] Gardner, W. and Martin, K. (**1995**), "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.* **97**(6), pp. 3907–3908. (Cited on page 31)

[37] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**), "DARPA TIMIT Acoustic-phonetic continuous speech corpus CD-ROM," *National Inst. Standards and Technol. (NIST)* . (Cited on page 15)

[38] Geier, M. and Spors, S. (**2012**), "Spatial Audio Reproduction with the SoundScape Renderer," in *27th Tonmeistertagung – VDT International Convention.* (Cited on page 88)

[39] Glasberg, B. R. and Moore, B. C. (**1990**), "Derivation of auditory filter shapes from notched-noise data," *Hearing research* **47**(1-2), pp. 103–138. (Cited on page 47)

[40] González, J. A., Peinado, A. M., Gómez, A. M., and Ma, N. (**2012**), "Log-spectral feature reconstruction based on an occlusion model for noise robust speech recognition," in *Proc. Interspeech*, pp. 2630–2633. (Cited on page 13)

[41] Hold, C. and Wierstorf, H. (**2016**), "Music mixes for comparison of wave field synthesis, surround, and stereo," URL `http://dx.doi.org/10.14279/depositonce-5173`. (Cited on page 76)

[42] Hold, C. and Wierstorf, H. (**2016**), "Signal feeds for creating the music mixes for comparison of wave field synthesis, surround, and stereo," URL `https://doi.org/10.5281/zenodo.55718`. (Cited on page 76)

[43] Hold, C., Wierstorf, H., and Raake, A. (**2016**), "Tonmischung fÃijr Stereophonie und Wellenfeldsynthese im Vergleich," in *Fortschritte der Akustik - DAGA 2015*, DEGA e.V., Aachen, Germany, pp. 1023–1026. (Cited on page 76)

[44] Hold, C., Wierstorf, H., and Raake, A. (**2016**), "The Difference Between Stereophony and Wave Field Synthesis in the Context of Popular Music," in *140th Conv. Audio Eng. Soc.*, p. 8. (Cited on page 76)

[45] Hosking, J. R. (**1990**), "L-moments: analysis and estimation of distributions using linear combinations of order statistics," *Journal of the Royal Statistical Society. Series B (Methodological)* , pp. 105–124. (Cited on page 47)

[46] Huang, X. and Weng, J. (**2002**), "Novelty and Reinforcement Learning in the Value System of Developmental Robots." in *Lund University Cognitive Studies*. (Cited on page 71)

[47] Ivaldi, S., Nguyen, S. M., Lyubova, N., Droniou, A., Padois, V., Filliat, D., Oudeyer, P. Y., and Sigaud, O. (**2014**), "Object learning through active exploration," *IEEE Transactions on Autonomous Mental Development* **6**, pp. 56–72. (Cited on page 71)

[48] Jack, C. E. and Thurlow, W. R. (**1973**), "Effects of degree of visual association and angle of displacement on the" ventriloquism" effect," *Perceptual and motor skills* **37**(3), pp. 967–979. (Cited on page 30)

[49] Jensen, K. and Andersen, T. H. (**2003**), "Real-time beat estimationusing feature extraction," in *International Symposium on Computer Music Modeling and Retrieval*, Springer, pp. 13–22. (Cited on page 47)

[50] Julier, S. J. and Uhlmann, J. K. (**2004**), "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE* **92**(3), pp. 401–422. (Cited on page 54)

[51] Kalinli, O. and Narayanan, S. (**2007**), "A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech," in *Interspeech*, pp. 1–4. (Cited on page 71)

[52] Klapuri, A. (**1999**), "Sound onset detection by applying psychoacoustic knowledge," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, IEEE, vol. 6, pp. 3089–3092. (Cited on page 47)

[53] Kohonen, T. (**1982**), "Self-organized formation of topologically correct feature maps," *Biological Cybernetics* **43**(1), pp. 59–69. (Cited on page 80)

[54] Lerch, A. (**2012**), *An introduction to audio content analysis: Applications in signal processing and music informatics*, John Wiley & Sons. (Cited on page 47)

[55] Liberman, M. C. (**1988**), "Response properties of cochlear efferent neurons: monaural vs. binaural stimulation and the effects of noise," *Journal of Neurophysiology* **60**(5), pp. 1779–1798, URL `http://jn.physiology.org/content/60/5/1779`. (Cited on page 9)

[56] Liu, J., Erwin, H., and Yang, G.-Z. (**2011**), "Attention driven computational model of the auditory midbrain for sound localization in reverberant environments," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pp. 1251–1258.

(Cited on page 11)

[57] Ma, N., Brown, G. J., and May, T. (**2015**), "Robust localisation of of multiple speakers exploiting deep neural networks and head movements," in *Proc. InterspeechâĂŹ15*. (Cited on page 89)

[58] Marquardt, D., Hadad, E., Gannor, S., and Doclo, S. (**2014**), "Optimal binaural LCMV beamformers for combined noise reduction and binaural cue preservation," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, https://pdfs.semanticscholar.org/95d1/d7d2459610c5031de650410c80452dc9fe2c.pdf. (Cited on page 44)

[59] May, T. and Dau, T. (**2014**), "Computational speech segregation based on an auditory-inspired modulation analysis," *The Journal of the Acoustical Society of America* **136**(6), pp. 3350–3359. (Cited on page 47)

[60] May, T., van de Par, S., and Kohlrausch, A. (**2012**), "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing* **20**(7), pp. 2016–2030. (Cited on page 47)

[61] Misra, H., Ikbal, S., Bourlard, H., and Hermansky, H. (**2004**), "Spectral entropy based feature for robust ASR," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, IEEE, vol. 1, pp. I–193. (Cited on page 47)

[62] Moritz, N., Anemüller, J., and Kollmeier, B. (**2011**), "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5492–5495. (Cited on page 47)

[63] Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (**2007**), "The mismatch negativity (MMN) in basic research of central auditory processing: a review." *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* **118**(12), pp. 2544–90, URL `http://www.ncbi.nlm.nih.gov/pubmed/17931964`. (Cited on page 72)

[64] Nguyen, S. M., Ivaldi, S., Lyubova, N., Droniou, A., Gerardeaux-Viret, D., Filliat, D., Padois, V., Sigaud, O., and Oudeyer, P. Y. (**2013**), "Learning to recognize objects through curiosity-driven manipulation with the iCub humanoid robot," *2013 IEEE 3rd Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL 2013 - Electronic Conference Proceedings* . (Cited on page 71)

[65] N.N. (**2014**), "Laboratory for Analysis and Architecture of Systems, Toulouse,

France," URL https://www.laas.fr/public/en/adream. (Cited on page 89)

[66] N.N. (**2016**), "MakeHuman^{TM}," URL http://www.makehuman.org/. (Cited on page 102)

[67] Patterson, R. D., Allerhand, M. H., and Giguère, C. (**1995**), "Time-domain modeling of periphal auditory processing: A modular architecture and software platform," *J. Acoust. Soc. Am.* **98**, pp. 1890–1894. (Cited on page 34)

[68] Patterson, R. D. and Holdsworth, J. (**1996**), "A functional model of neural activity patterns and auditory images," *Advances in speech, hearing and language processing* **3**(Part B), pp. 547–563. (Cited on page 47)

[69] Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (**2011**), "The timbre toolbox: Extracting audio descriptors from musical signals," *The Journal of the Acoustical Society of America* **130**(5), pp. 2902–2916. (Cited on page 47)

[70] Portello, A., Bustamante, G., Danès, P., Piat, J., and Manhès, J. (**2014**), "Active Localization of an Intermittent Sound Source from a Moving Binaural Sensor," in *Forum Acustium (FA'2014)*, Krakow, Poland. (Cited on page 52)

[71] Portello, A., Danès, P., Argentieri, S., and Pledel, S. (**2013**), "HRTF-Based Source Azimuth Estimation and Activity Detection from a Binaural Sensor," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'2013)*, Tokyo, Japan. (Cited on page 52)

[72] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (**2011**), "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. (Cited on page 78)

[73] Premakumar, P. (**2016**), "A* (A Star) search path planning tutorial," URL https://de.mathworks.com/matlabcentral/fileexchange/26248-a---a-star--search-for-path-planning-tutorial?requestedDomain=www.mathworks.com. (Cited on page 93)

[74] Qian, J., Hastie, T., Friedman, J., Tibshirani, R., and Simon, N. (http://gts.sourceforge.net/ **2013**), "Glmnet for Matlab," http://gts.sourceforge.net/. (Cited on page 47)

[75] Rennie, S. J., Hershey, J. R., and Olsen, P. A. (**2010**), "Single-channel multitalker speech recognition," *IEEE Signal Process. Mag.* **27**, pp. 66–80. (Cited on pages 12 and 13)

[76] Roman, N., Wang, D., and Brown, G. J. (**2003**), "Speech segregation based on sound

localization," *The Journal of the Acoustical Society of America* **114**(4), pp. 2236–2252. (Cited on page 38)

[77] Roy, N., Mccallum, A., and Com, M. W. (**2001**), "Toward Optimal Active Learning through Monte Carlo Estimation of Error Reduction," in *international conference on Machine Learning*. (Cited on page 72)

[78] Ruesch, J., Lopes, M., Bernardino, A., Hörnstein, J., Santos-Victor, J., and Pfeifer, R. (**2008**), "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub," *Proceedings - IEEE International Conference on Robotics and Automation* , pp. 962–967. (Cited on page 71)

[79] Ryan, R. M. and Deci, E. L. (**2000**), "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions." *Contemporary Educational Psychology* **25**(1), pp. 54–67, URL `http://www.ncbi.nlm.nih.gov/pubmed/10620381`. (Cited on page 71)

[80] Shannon, C. E. (**1948**), "A Mathematical Theory of Communication," *Bell System Technical Journal* **27**, pp. 379–423, 623–656. (Cited on page 65)

[81] Spence, G. C. and Driver, J. (**1994**), "Covert spatial orienting in audition: exogenous and endogenous mechanisms," *Journal of Experimental Psychology* **20**, pp. 555–574. (Cited on page 11)

[82] Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (**1988**), "Lateralization of complex binaural stimuli: A weighted-image model," *J. Acoust. Soc. Am.* **84**, pp. 156–165. (Cited on pages)

[83] Sutton, R. S. and Barto, A. G. (**1998**), *Reinforcement Learning: An Introduction*, The MIT Press. (Cited on page 98)

[84] Teret, E., Pastore, M., and Braasch, J. (**2016**), "The iniĭňĆuence of signal type on the internal auditory representation of a room," *J. Acoust. Soc. Am.* **tbd.**, pp. (submitted). (Cited on page 24)

[85] Tibshirani, R. (**1996**), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)* , pp. 267–288. (Cited on page 47)

[86] Traa, J. (**2013**), "Least-Squares Intersection of Lines," URL `http://cal.cs.illinois.edu/~johannes/research/LS_line_intersect.pdf`. (Cited on page 94)

[87] Tzanetakis, G. and Cook, P. (**2002**), "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing* **10**(5), pp. 293–302. (Cited on page 47)

[88] v. d. Malsburg, C. (**1999**), "The What and Why of Binding: The Modeler's Perspective," *Neuron* **24**, pp. 95–104. (Cited on page 91)

[89] Varga, A. and Moore, R. (**1990**), "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 845–848. (Cited on page 12)

[90] Walther, T. and Cohen-L'hyver, B. (**2014**), "Multimodal feedback in auditory-based active scene exploration," in *Proc. Forum Acusticum*, Kraków, Poland. (Cited on page 63)

[91] Wang, D. L. and Brown, G. J. (Eds.) (**2006**), *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley/IEEE Press. (Cited on page 13)

[92] Wierstorf, H., Geier, M., and Spors, S. (**2011**), "A Free Database of Head Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances," in *Audio Engineering Society Convention 130*, URL `http://www.aes.org/e-lib/browse.cfm?elib=16564`. (Cited on page 45)

[93] WillowGarage (**2014**), "Open Source Computer Vision Library," URL `http://sourceforge.net/projects/opencvlibrary/`. (Cited on page 107)

[94] Winkowski, D. E. and Knudsen, E. I. (**2006**), "Top-down gain control of the auditory space map by gaze control circuitry in the barn owl." *Nature* **439**(7074), pp. 336–339. (Cited on page 11)

[95] Woodruff, J. and Wang, D. (**2013**), "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE T. Audio. Speech.* **21**(4), pp. 806–815. (Cited on page 11)

[96] Woodruff, J. and Wang, D. L. (**2012**), "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.* **20**(5), pp. 1503–1512. (Cited on page 11)

[97] Zeiler, S., Meutzner, H., Abdelaziz, A. H., and Kolossa, D. (**2016**), "Introducing the Turbo-Twin-HMM for Audio-Visual Speech Enhancement," in *Proc. Interspeech.* (Cited on page 78)