Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Audiovisual Scene Analysis in Online-Adaptation of Audio-Motor Maps

Rujiao Yan, Tobias Rodemann, Britta Wrede

2013

Preprint:

This is an accepted article published in Transactions on Autonomous Mental Development. The final authenticated version is available online at: https://doi.org/[DOI not available]

Computational Audiovisual Scene Analysis in Online Adaptation of Audio-Motor Maps

Rujiao Yan, Tobias Rodemann, and Britta Wrede

***Abstract*—For sound localization, the binaural auditory system of a robot needs audio-motor maps which represent the relationship between certain audio features and the position of the sound source. This mapping is normally learned during an offline calibration in controlled environments, but we show that using Computational AudioVisual Scene Analysis (CAVSA), it can be adapted online in free interaction with a number of a-priori unknown speakers. CAVSA enables a robot to understand dynamic dialog scenarios, such as the number and position of speakers, as well as who is the current speaker. Our system does not require specific motions of the robot and thus can work during other tasks. The performance of online-adapted maps is continuously monitored by computing the difference between online-adapted and offline-calibrated maps and also comparing sound localization results with ground truth data (if available). We show that our approach is more robust in multi-person scenarios than the state of the art in terms of learning progress. We also show that our system is able to bootstrap with a randomized audio-motor map and adapt to hardware modifications that induce a change in audio-motor maps.**

I. Introduction

SOUND plays an important role in human communication and perception of the surrounding scene. Humans determine where a sound came from by using only two ears. Inspired by the human sound localization, our humanoid robot head is equipped with two microphones embedded at the positions of the ears [1]. From the difference in the recorded signal between the left and right ear, the position of the sound source can be inferred. The two most widely used cues are the Interaural Intensity Difference (IID) and the Interaural Time Difference (ITD) [2]. IID is the difference in the sound amplitude between the left and right ear (microphone). It makes use of the head-shadow effect, that the head absorbs some of the sound energy and attenuates the sound signal arriving at the far ear especially at higher frequencies. Hence IID is influenced by the size, form, material and density of the robot head. Furthermore, IID depends on azimuth and to some extent on elevation and distance of the sound source relative to the head as well as the frequency spectrum of the sound. ITD, the delay in arrival time of a sound signal between two ears (microphones), has similar characteristics as IID but in its form reflects more the dimension and shape of the head than its material. ITD calculated from interaural phase difference is ambiguous at high frequencies, i.e. many different source positions would generate the same ITD value. In contrast, IID is only significant at high frequencies, since low frequency signals are not attenuated very much by the head. Therefore, we need both IID and ITD cues to perform sound localization in the full range of frequencies. In addition to IID and ITD there are other possible cues for sound source localization, such as spectral cues [3]–[6], which are implemented on our robot but not used in this work.

Since both cues show a strong dependency on the sound source position, it is possible to derive this position from a measurement of IID and ITD. For this purpose, the system needs to know the relation between sound source position in motor coordinates (azimuth and elevation angle) and typical values of IID and ITD. This relation is called **audio-motor map**, which is closely related to the concept of Head Related Transfer Functions (HRTFs) [7]. We concentrate on audio-motor maps for azimuth to ease the description, because it is sufficient for many practical applications to compute only the azimuth angle and to ignore the sound's elevation and distance. Especially in this work, we deal with human speakers who tend to have approximately the same height (e.g. few deviations in elevation) and roughly the same distance. Note that our system is principally capable of estimating also elevation and distance using only two ears [1], [8], but this ability is not used in the described system. This is basically for simplicity, as it would require a substantially larger audio-motor map and more processing in the audio and visual domain.

How to obtain the relation between binaural cues and azimuth angles? For a perfectly spherical robot head of diameter d with two microphones at exactly opposite positions there is an analytical approximation to ITD for azimuth angle (θ) [2]:

$$ITD(\theta) = \frac{d}{2v} \cdot (\theta + \sin(\theta)) \approx \frac{d}{v} \cdot \sin(\theta) \qquad , \qquad (1)$$

with v as the speed of sound in air. This equation is only approximately valid because it shows no frequency dependency. To overcome this disadvantage, Viste et al. [7] added a scaling factor α_f to Eq. 1, that is measured for the individual head and each frequency f. They also approximated IID using such a scaling factor β_f as follows:

$$IID(\theta, f) = \beta_f \cdot \frac{\sin(\theta)}{v}. \qquad (2)$$

However, it was reported that, the models of ITD and IID are less accurate than HRFTs, especially in high frequencies. One of the possible reasons is that these models are based on the assumption of a perfectly spherical head. Therefore,

R. Yan and B. Wrede are with the Research Institute for Cognition and Robotics, Bielefeld University, 33594 Bielefeld, Germany (e-mail: ryan@cor-lab.uni-bielefeld.de; bwrede@cor-lab.uni-bielefeld.de)

T. Rodemann is with the Honda Research Institute Europe, Carl-Legien-Street 30, 63073 Offenbach/Main, Germany (e-mail: tobias.rodemann@honda-ri.de)

audio-motor maps have to be learned for more general robot architectures.

In the following we will introduce drawbacks of the related work in learning of audio-motor maps (motivation of using CAVSA), what is CAVSA and the aim of our approach.

A. Related Work in Learning of Audio-Motor Maps

Audio-motor maps are normally learned during offline calibration in controlled environments. In a typical calibration scenario, a number of sounds are played from a set of defined positions. For each position, localization cues are computed, averaged and stored in an audio-motor map. Weng et. al. [9] train such a map for 3D sound localization using 4 microphones. Willert et. al. [10] use a neural network trained in a similar fashion instead of using an audio-motor map. The drawback of an offline calibration is that it is a costly process that needs to be run in a controlled environment, so that the robot has to be put out of operation and recalibrated by an expert in the field using a special set-up. Furthermore, audio-motor maps are highly variable and often change due to different room environments and hardware modifications such as changes of microphone type, modification of robot head and microphone position. Therefore, from an application point of view it is highly beneficial to develop an unsupervised, online adaptation of the audio-motor maps so that the system always works at maximum efficiency. It has been proven that the brain of humans and animals also has to learn such an audio-motor map in interaction with the environment and that this process is active even after infancy [11]–[17]. For humans, the audio-motor map is changing during the full growth process of the head and external ears, therefore a continuous update must be possible. An update of the audio-motor map is conceptually simple in engineering applications. All that is needed is the measurement of IID/ITD and the true position of the sound source. Since the measurement of binaural cues is easy, the problem boils down to the question how to determine the sound source position when one can not trust the sound localization system.

In a previous work [18], our group has shown that a robot head which can execute pan motions is able to learn the audio-motor map using only audio input in a simple environment. Unfortunately, this approach is rather slow, requires permanent head motions, is only applicable to estimate the azimuth angle, and just works under the assumption that only one sound source exists in the environment. It might however serve as a starting point for bootstrapping the initial audio-motor map. Inspired by the important role of vision for the calibration of audio-motor maps in animals and humans [19]–[21], several recent papers [22], [23] suggest to use visual feedback to provide the sound source position. These approaches are very appealing but are demonstrated only in simple scenarios. Nakashima et al. [23] attached a red color marker on the loudspeaker to indicate the source, while Hörnstein [22] et al. used a face tracker. Both approaches are shown to be very successful in their scenarios but would fail under less constrained conditions. In the case of Nakashima, any additional sound source or any other red object would make a correct link of visual and auditory information unreliable. For Hörnstein's system, more than one person or the wrong speaker in the camera image would lead to confusion and potentially wrong updates of the mapping. The underlying problem in these approaches is that they use a very simple scene representation consisting of only one visual object and its position. Hence, a correct match can not be made for scenarios with several visual objects. Besides, both methods need extra head motions to search for the visual correspondence of the sound source, so that the robot may have to interrupt its normal operation for a certain period of time. In this work we will demonstrate that CAVSA, as introduced below, is able to deal with more complex scenarios and better adapt the audio-motor map. Our system does not require specific motions of the robot, so that audio-motor maps can be continuously adapted online during the normal operation of the robot.

B. Computational AudioVisual Scene Analysis

We apply Computational AudioVisual Scene Analysis (CAVSA) to search for the visual part of the current sound source for online adaptation of audio-motor maps. In general, CAVSA enables a robot to understand its surroundings when interacting with humans. For example, the robot needs to know how many speakers are present, where the speakers are and who is currently talking to the robot. In the application specified in this article, CAVSA is used to link an utterance to a face when we hear an utterance and see many faces.

In the field of robot audition the topic of building an auditory scene analysis (ASA) has been studied since quite some time. Recent books on Computational Auditory Scene Analysis (CASA) describe intensive work in this field [24]–[26]. However, the focus of most current work is on the processing of several concurrent speakers, i.e. separating sound sources in a mixed recording and performing a speech-to-text conversion on the separated signals. While this is an important step, the issue of how to group different utterances from different speakers in natural environments is often neglected. The CAVSA proposed in this article intentionally ignores the issue of sound source separation by assuming that all speakers talk in sequence (or sound source separation has been implemented) and that background noise is stationary, and focuses on the analysis of a dialog scenario with an unknown number of speakers who have not been seen or heard before.

Inspired from infant learning we employ simple auditory and visual features to represent speakers. As shown in Fig 1, features are first collected in a compressed form called proto-objects. Proto-objects, originally a psychophysical concept, are segmented perceptual units and describe "objects" before they are recognized. Therefore, proto-objects are a mid-level scene representation between pixel (or voxel) and semantic object level [27]–[29]. Proto-objects can be described as "something" coherent at a certain position (e.g. in the robot's camera image) or at a certain point in time (e.g. for the auditory signal). The process that defines proto-objects first segments regions (x/y in the camera image, time/frequency in audio signals) based on homogeneity, edges or object detection, calculates a reference point (e.g. the center of segment in the image or the starting

point of an utterance) and then computes a number of simple cues inside the segment, like color histogram in an image region or the mean signal energy of an utterance. In this work we use only a small set of simple features that are sufficient for online adaptation of audio-motor maps, for example positions of people in visual proto-objects and binaural cues (IID and ITD) in audio proto-objects. It is beneficial to apply such proto-objects. First, proto-objects can be easily extended with new features to improve performance without changing the architecture of the system. Second, since proto-objects are sparse so that there are only a few visual proto-objects per image frame and one audio proto-object every few seconds, maintaining this representation comes at a low computational cost. Third, proto-objects can be tracked, pointed or referred to without identification and enable a flexible interface to behavior-control in robotics.

Then, both audio and visual proto-objects are stored in two separate short-term memories (STMs). New proto-objects are integrated into the STM by comparing them with those already stored in memory. Those similar to existing ones will be merged with the most similar proto-objects in memory, while those dissimilar to all existing ones will be added to the STM. The comparison of proto-objects is based on one or multiple simple auditory or visual features. Proto-objects that are not updated for more than a certain period are removed. In this manner, the STM can memorize persons who disappear from the camera field of view for a while.

Finally, audio and visual proto-objects are integrated based on either temporal or temporal and spatial coincidence, depending on the number of visual proto-objects. It has to be mentioned that audiovisual integration consists of two steps: audiovisual association and cue integration. The first step is to link an auditory and a visual signal that are caused by the same event. The second step is to integrate multisensory cues, that are auditory and visual position in this context. We focus on the first step and simplify the second step by considering the visual position as the position of the speaker, since visual position estimation is normally more accurate than auditory. Integration of auditory and visual position signals can be found e.g. in [30].

Temporal coincidence (synchrony) has been identified as one of the most important factors determining whether or not multisensory integration takes place [31]. If only one auditory and one visual stimulus are temporally coincident, they are perceptually coherent, even when they are spatially disparate, such as in the ventriloquism effect [32]. If more than one source exists, we need other information such as position information to avoid ambiguity. The closer a visual stimulus is to an auditory stimulus, the more probable they are perceived as having a common cause [33]. Actually, it is known that prior spatial correlation between auditory and visual stimuli is not required for audiovisual integration in baby cats and young barn owls [34], [35]. When the animals are raised in artificial environments where auditory and visual stimuli are temporally coupled but spatially not coherent, multisensory neurons in the superior colliculus (SC) are also able to integrate these stimuli. Spatial coincidence appears to be learned early in life. The learning itself is an online process to optimally adapt to the

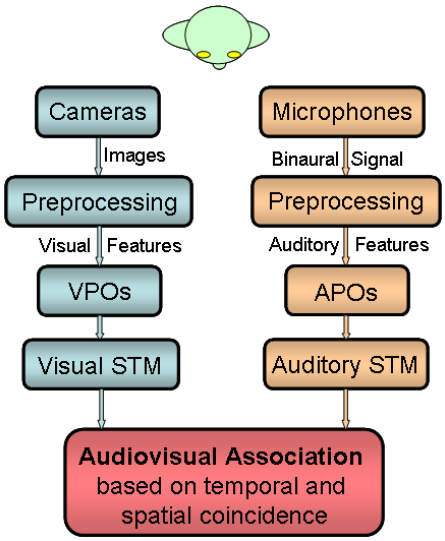

Fig. 1. System architecture of Computational AudioVisual Scene Analysis (CAVSA). APO: audio proto-object, VPO: visual proto-object, STM: short-term memory.

environment. Inspired by this, our audiovisual integration is purely based on temporal coincidence when there is only one auditory and one visual stimulus (one utterance and one face in the scene). Otherwise, when more than one face appears in the scene, audiovisual integration is based on both temporal and spatial coincidence.

Audiovisual association is an important issue for CAVSA. A large volume of work has been done relating to how auditory and visual signals are linked in multi-person environments for different applications. The first group focuses on temporal coincidence between speech and visual activity (lip motion or body motion) [36]–[41]. These methods need only one camera and one microphone, and they do not make a prior measurement of auditory and visual position. Their main application is speaker indexing in multimedia data for speaker diarization (annotation of "who spoke when") or as preprocessing for multimodal person identification. Most of them can only work in constrained environments to avoid artifacts, e.g. non-talking persons are not allowed to move significantly. Therefore, these methods usually require careful tuning of quantities such as camera image resolution and temporal resolution. The second group integrates auditory and visual signals based on both temporal and spatial coincidence [42]–[46]. These methods do not need high temporal resolution in contrast to the first group, but focus on position information. They are usually proposed for multi-person tracking. For this purpose, they use binaural microphones or microphone arrays to localize sound, and estimate 3D visual positions using stereo cameras or even more cameras in a smart-room environment. While the robot head is moving, a speaker may leave the field of view for a while and reappear again, as most robots have a limited field of view. In this case, these methods do not memorize the speaker and would consider him as a new person. In comparison to these methods, our system is based on either temporal or temporal and spatial coincidence depending on the number of visual sources, and can memorize a person who disappears

briefly by using the STM concept. Additionally, the current sound is sometimes related to a person who has not yet been seen. We consider this situation as well by applying a view memory that memorizes how long a certain position has not been looked at.

C. Approach

Our approach aims at online adaptation of audio-motor maps using CAVSA during the normal operation of the robot. In our approach, CAVSA applies the current audio-motor maps for sound localization and in turn supports online adaptation of audio-motor maps by searching for the corresponding visual part of the current sound in multi-person scenarios. On its own, the interaction of high-level scene representations and low-level sensory-motor maps is an interesting field of study, since most approaches [22], [23], [47], [48] that study the development of those maps rely on very simplified sensory environments that are often unrealistic and are not able to explain how animals or humans manage to initially learn and continuously adapt these maps.

Our previous work [49] introduced an earlier version of CAVSA that was applied in online adaptation of audio-motor maps. Compared with prior online adaptation methods [22], [23], our approach was shown to be more robust in scenarios with more than one speaker and when speakers dynamically enter and leave the scene. A later version [50] considered the situation that the current sound is sometimes related to a person who has not yet been seen. In this work, we focus on the issue of the bootstrapping and adaptation ability. We were interested to see if the system is able to learn the audio-motor maps from a random initialization in different scenarios and how it responds to changes of the mapping during operation, when acoustic conditions such as the shape of the robot head or microphones are modified. To measure the map quality, we compute the difference between online-adapted and offline-calibrated audio-motor maps and also compare sound localization results with ground truth data (if available).

Our system uses a humanoid robot head with a pair of cameras and a pair of microphones. All experiments are conducted in a normal office room of size $4x3x2.8m$ with typical echo and background noise conditions such as noise from computers and air conditioner, door shut noise from neighboring office rooms. As sound sources, we use pre-recorded utterances played from a loudspeaker and real human speakers.

The rest of the article is organized as follows. Section II introduces how to process auditory and visual signals, in order to achieve the CAVSA. Section III describes the application of CAVSA in online adaptation of audio-motor maps. Section IV presents the experiments and results. Finally, discussion and future issues are given in Section V.

II. Scene Representation

Low level sensory representations are often unsuited to guide behavior in real-world environments, due to rapid changes in the signals, measurement noise and a high complexity of scenes. It is therefore necessary to build a more stable and robust representation of the environment - a scene representation.

For computational audiovisual scene analysis, our system first collects auditory and visual features in a compressed form called proto-objects, that is designed for operation in realistic environments. Proto-objects for the same speaker are then grouped together in auditory and visual Short-Term Memories (STM), respectively. Finally, visual and audio proto-objects of the same speaker are integrated. Fig. 1 schematically illustrates the system architecture of the CAVSA. Proto-objects, auditory and visual processing, the concept of STM as well as audiovisual integration are described below.

A. Proto-Objects

Most previous approaches such as [51] rely on low-level position estimations, i.e. the position of a face in the current image or the computed sound position in the certain time frame. These measurements are often very noisy and dealing with missing information (e.g. no new sound localization when the speaker is silent) is difficult. Therefore, our approach uses an intermediate representation of audio and visual sensory information: proto-objects. In the past we have used both visual [28] and audio [29] proto-objects separately. In this work we first combine the two approaches into one consistent framework. The term "proto-object" is inspired by earlier work of [36], [52], [53] on the representation of sensory entities in the brain. In contrast to objects, as for example in object recognition, proto-objects do not need to refer to a semantically meaningful unit. Rather they represent "something" coherent in the environment. In the visual domain, this could be a number of neighboring pixels within a certain distance range or pixels of a certain color [54]. We segment visual regions based on face detection, which is a low-level mechanism. It was found that human newborns have already inborn face detection mechanisms [55], [56]. In the audio domain, we use a simple threshold-based segmentation process similar to Voice Activity Detection (VAD) to separate sound pieces from the background [57]. Proto-objects can be used to represent and address behaviorally relevant objects, object parts or group of objects in the environment.

A proto-object can combine an arbitrary number of features, and new features can be added easily. A typical visual proto-object would contain a position (for example center of gravity) and a compact representation of one or more visual features such as the average color in the segment. Consecutive frames would generate proto-objects of very similar characteristics (e.g. position and color). Therefore it is easy to track these proto-objects over time [54]. In the auditory domain, proto-objects typically relate to human utterances, from single words to whole sentences. There are often periods of silence between two proto-objects, so that continuous tracking of the sound source is impossible. Audio proto-objects can contain information about the timing of the sound (such as the start time), the length and averaged audio features (such as the mean estimated position over the complete length of the sound, histograms of

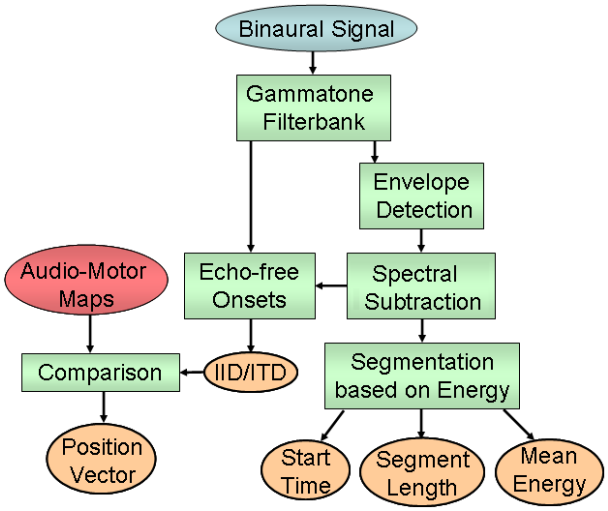

Fig. 2. Audio processing

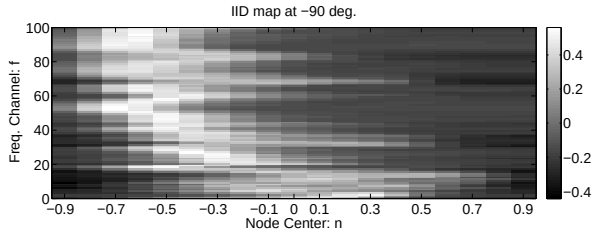

Fig. 3. An example of the IID map at -90°, corresponding to the population code vectors of measured IID we would expect if a sound source were at position $p = -90^\circ$. The vector in each frequency channel is normalized to mean 0 and norm 1.

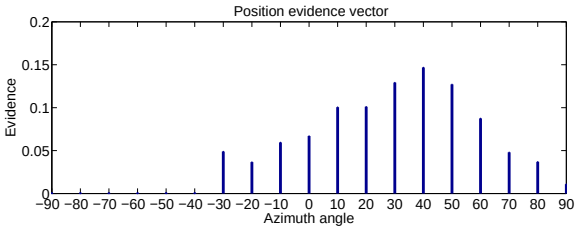

Fig. 4. An example of a position evidence vector. The vector is normalized to sum 1. Here the estimated azimuth is 40°.

localization features like IID and ITD, and the mean amplitude of the sound). In this work we focus on a small set of simple features that are sufficient for online adaptation of audio-motor maps. In the current context, positions of people (faces) and the time of the current frame are stored in visual proto-objects, while audio proto-objects contain start time and length of sound segments, mean energy, estimated position and binaural cues (IID and ITD). In Section II-B and II-C it is explained how to obtain these auditory and visual features, respectively.

B. Audio Processing

We employ a biological-inspired audio processing system using two microphones on a humanoid robot head. The microphones are DPA Miniature Microphones 4060-BM, located roughly at the position of human ears and surrounded by pinnae-like structures [1]. As shown in Fig. 2, the first stage in audio-processing is a Gammatone Filterbank (GFB), which is considered to be a good model of the human cochlea [58]. The GFB transforms the signal from the temporal into the spectro-temporal domain, using a number (100) of bandpass filters with logarithmically increasing center frequencies. In a further preprocessing step we extract the signal envelope for every frequency band. In order to deal with stationary background noise we employ a spectral-subtraction technique that estimates the level of noise in each frequency channel and then reduces this noise. This method is very effective against noise sources like air-conditioning or computer fan noise. Thereafter, signal energy is calculated as the sum over envelope signals in all frequency channels. Based on the energy, audio streams are segmented. An audio segment begins when the signal energy exceeds a threshold and ends when the energy falls below this threshold. Next, we derive start time, length and the mean energy over all samples of the segment, and save these features in an audio proto-object.

The most important audio feature to be computed is the position of a sound source. A typical problem in sound localization is the negative impact of echoes. We use a biologically inspired strategy, the precedence effect, to reduce the impact of echoes [59]. The basic principle is the detection of signal onsets, which are quick and strong changes in the frequency channel's envelope. We only measure audio cues for sound localization during the signal onsets (before echoes arrive). We have shown in several previous papers that our system can efficiently operate even in noisy and echoic environments [1], [59], [60]. Since audio cues, IID and ITD, depend on the frequency spectrum, they are measured for each frequency channel f. The measurement of audio cues is explained in [59]. IID values are represented as a ratio in the range from $(-1,1)$, while the ITD values are in the range $[-0.9,0.9]ms$. Audio cues $c^i(f)$ ($i=1$ for IID, $i=2$ for ITD) in each frequency channel f are then re-encoded into a population code. In a population coding, feature values are indirectly represented by the activity of a number of nodes (neurons) [61]. For the encoding of audio cues, we use a set of 19 nodes with response centers at $(-0.9,-0.8,...,0,...,0.8,0.9)$ that respond to cue values. Every IID or ITD value leads to an activation in the nearest nodes. The population code vector of $c^i(f)$ for frequency channel f is defined by a Gaussian function:

$$C^i(f,n) = \exp\left(\frac{-(c^i(f)-n)^2}{2\cdot\delta^2}\right), \qquad (3)$$

where n stands for node center and the standard deviation δ is set to 0.1. Fig. 3 shows for example the expected population code vectors of measured IID $C^1(f,n)$ when a sound source were at -90°. All measurements are added over time for an audio proto-object and the sum is denoted as $C'^i(f,n)$. Given audio-motor maps $M_p^i(f,n)$ that will be explained in Section III, we compare the population response $C'^i(f,n)$ with $M_p^i(f,n)$ for all positions p by computing scalar products to acquire the position evidence vector W_p^i. The peak in the position vector W_p ($W_p = \sum_{i=1}^{2} W_p^i$) is taken as the estimated sound source position. Fig. 4 shows an example of a position evidence vector with the peak at 40°. In this

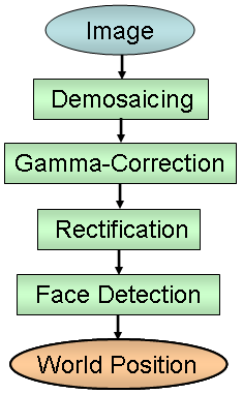

Fig. 5. Visual processing

work we save also population-coded cues $C'^{i}(f,n)$ and the position evidence vector W_p in an audio proto-object besides start time, length and the mean energy. A filtering of audio proto-objects based on segment length and energy is also performed, since background sounds are rather short and have less energy than robot directed calls. Filtering of background sounds is explained in [62].

C. Visual Processing

Our robot head is equipped with a set of stereo color cameras (Matrix Vision BlueFOX 224C). The camera has a horizontal field of view of 90°. Since we do not rely on visual distance estimation in this work, only a single camera is used. Note however that our group has used visual distance estimation in human-robot interaction before [54]. As shown in Fig 5, a "demosaicing" stage follows after image acquisition to reconstruct a full-color image. Then gamma correction and rectification are performed to compensate for properties of human vision and correct image distortion, respectively. After that, we extract visual features of a person to form a visual proto-object. In the camera field of view, the presence of persons is detected by searching for their faces. We assume that persons talking to the robot look at its face most of the time. A frontal face detection algorithm developed by Viola and Jones [63] is chosen for its high speed and high detection rates. For each of these proto-objects, the center and size of each face in the camera image is computed. We use only one camera and keep the distance between robot head and speakers about $1m$. Using the distance information and saccade maps (see [64]), we convert camera positions into 3D world coordinates, and store them as well as the time of the current image in the visual proto-object.

D. Short-Term Memory

In biology, a short-term memory [65] contains the most recently perceived items and actions used for a certain time. We use a similar concept for our scene representation. Initially the short-term memory (STM) is empty. When a new audio or visual proto-object appears, it is copied into the STM. Unless it is updated, it will disappear after a certain time-out T. The larger the parameter T, the longer a proto-object is stored in STM after the corresponding person disappears. If T is too small, the STM can not remember a person who leaves for a while. Conversely, if T is too large, STM may contain too many proto-objects and is not able to represent the current scene. We set $T = 100s$ in our experiment.

We determine if an incoming proto-object is new or corresponds to a proto-object in the STM, based on the distance. The assumption is that people do not change their position much in a dialog scenario, so that the position distance of the same person is small between two adjacent views. When an incoming proto-object appears at position p_{in}, it will be compared with each already in the STM. We denote the position of a proto-object in STM as p_n $(n \in [1,N])$, where N is the number of proto-objects in STM. The incoming proto-object is merged with the proto-object at position p_k $(k \in [1,N])$, if

$$k = \arg\min_n d(p_{in}, p_n) \wedge d(p_{in}, p_k) < \Theta_s, \tag{4}$$

where $d(p_{in}, p_n)$ describes the pairwise position distance and Θ_s is the threshold of position distance. $d(p_{in}, p_n)$ can be the difference of azimuth angle in spherical coordinates (for auditory or visual signals) or Euclidean difference in 3D world coordinates (for visual signals). Then a merge operation is done by a low pass filter:

$$p_k(t) = p_k(t-1) \cdot w + p_{in}(t) \cdot (1-w), \tag{5}$$

Here t is time step and the weight w is defined as an activity function:

$$w = e^{-\frac{t_k}{\tau}}, \tag{6}$$

where τ is a decay factor and t_k stands for the time in seconds since the proto-object at p_k has been updated the last time. If the proto-object is currently updated, t_k is set to 0. w indicates the reliability of the position $p_k(t-1)$ that is updated in the recent time step. The longer a proto-object has not been seen, the less reliable its features become. τ is set to $5s$ based on empirical test. Otherwise, if no similar proto-object is found, the incoming proto-object is added to the STM as a new entry. Theoretically, the number of proto-objects in the STM can become very large, but due to the sparsity of proto-objects (only few are generated in every image frame, and there is only one audio proto-object at a time) and the time-out T, the effective number is typically very low.

In the case of visual STM, we notice that when more than one person appears in the camera image and more than one proto-object is stored in STM, two or more incoming proto-objects may tend to select the same proto-object in STM using the rule in equation (4). To avoid the selection confusion, we use a simple greedy mapping method. Let X_m denote the m-th current incoming proto-object, Y_n $(n \in [1,N])$ denote the n-th proto-object in STM, where M represents the number of current incoming proto-objects and N the number of proto-objects in STM. If d_{mn} stands for the position distance between X_m and Y_n, all distances d_{mn} build a matrix D' with M rows and N columns. For the greedy algorithm, the pair of proto-objects with the minimum distance is first selected. If the distance is smaller than Θ_s, the two proto-objects are merged. Then their distance related row and column are eliminated from distance

matrix D'. In the next step, the next nearest pair of proto-objects is selected. The process is repeated until all incoming proto-objects are assigned or the minimum distance is larger than Θ_s. In the latter case, the remaining incoming proto-objects, which are not yet assigned, are stored in STM as new entries. A similar usage of such a greedy algorithm can be found in [37].

E. Audiovisual Integration

In this section we explain how audio and visual proto-objects are linked using their position information.

Auditory position evidence vectors and visual positions in world coordinates must be converted to the same metric, we choose motor coordinates (azimuth and elevation). In this paper we concentrate only on azimuth as mentioned. The azimuth angle of an audio proto-object is taken as the peak position in its position evidence vector, while the azimuth angle of a visual proto-object is inferred from 3D world position.

Let p_{a_i} and p_{v_j} denote the azimuth angle of audio proto-object A_i ($i \in [1, N_A]$) and visual proto-object V_j ($j \in [1, N_V]$), respectively. N_A stands for the number of audio proto-objects in auditory STM, and N_V the number of visual proto-objects in visual STM. The relative probability that audio proto-object A_i and visual proto-object V_j have a common cause can then be approximated as:

$$P_{common}(p_{a_i}, p_{v_j}) = \exp\left(\frac{-|p_{a_i} - p_{v_j}|^2}{2 \cdot \delta_{AV}^2}\right), \tag{7}$$

where the standard deviation δ_{AV} is computed as the mean absolute difference in estimated azimuth between an audio and a visual proto-object which belong to the same speaker. The calculation of δ_{AV} is explained later in Section III-A.

We also consider the situation that the audio proto-object A_i is sometimes related to a visual proto-object V_{N+1} that has not yet been seen [50]. We set a view memory $U(p)$, in which each element at position p is between 0 and 1. The larger the value $U(p)$ is, the longer the time since position p has been last attended. $U(p)$ is described as

$$U(p) = 1 - e^{-\frac{t_p}{T_u}}, \tag{8}$$

where t_p indicates the time in seconds since position p has been viewed the last time. The initial value of t_p is set to ∞, so that $U(p) = 1$ if position p has never been attended. If $t_p = 0$, then $U(p) = 0$ which means that position p is currently in the field of view. T_u is a time constant to decay the activity and it is set to $100s$ to match the parameter T for STM.

By means of the view memory $U(p)$, the probability that A_i is associated with V_{N+1} is described by:

$$P_{common}(p_{a_i}, p_{v_{N+1}}) = \frac{\sum_p U(p) \cdot P_{common}(p_{a_i}, p)}{\sum_p P_{common}(p_{a_i}, p)}. \tag{9}$$

Then we search for the visual proto-object V_l with the maximum probability P_{common}:

$$l = \underset{j \in [1, N_V+1]}{\arg\max}\ P_{common}(p_{a_i}, p_{v_j}) \tag{10}$$

If $V_l = V_{N+1}$, the matched visual proto-object has very probably not yet been seen, and the audiovisual integration can not be executed. Otherwise, we check how reliable the integration of A_i and V_l is using the entropy of the set of normalized probabilities [50]. If one visual proto-object shows a very high probability and all other visual proto-objects have a low probability, this indicates a reliable association. Conversely, when all visual proto-objects have quasi equal probability, the association is unreliable. The normalized probability is computed as

$$\widehat{P}_{common}(p_{a_i}, p_{v_j}) = \frac{P_{common}(p_{a_i}, p_{v_j})}{\sum_{i=1}^{N_V+1} P_{common}(p_{a_i}, p_{v_j})}, \tag{11}$$

and the entropy for A_i is given as

$$H_i = \frac{-\sum_{j=1}^{N_V+1} \widehat{P}_{common}(p_{a_i}, p_{v_j}) \cdot \log_2 \widehat{P}_{common}(p_{a_i}, p_{v_j})}{\log_2(N_V + 1)} \tag{12}$$

Here, the division by $\log_2(N_V + 1)$ ensures that the maximal H_i is 1 to easily set a threshold Θ_H. If entropy H_i is larger than Θ_H, A_i and V_l are not associated. Θ_H is set to 0.8 based on empirical test. The method is summarized as Algorithm 1.

Algorithm 1 Audiovisual integration

for $j = 1$ to N_V **do**
  Calculate $P_{common}(p_{a_i}, p_{v_j})$ based on Eq. 7
end for
Calculate $P_{common}(p_{a_i}, p_{v_{N+1}})$ as in Eq. 9
Search for V_l that satisfies

$$l = \underset{j \in [1, N_V+1]}{\arg\max}\ P_{common}(p_{a_i}, p_{v_j}) \wedge V_l \neq V_{N+1}$$

if $\exists\ V_l$ **then**
  for $j = 1$ to $N_V + 1$ **do**
    Calculate $\widehat{P}_{common}(p_{a_i}, p_{v_j})$ based on Eq. 11
  end for
  Calculate entropy H_i as in Eq. 12
  if $H_i \leqslant \Theta_H$ **then**
    Integrate A_i with V_l
  end if
end if

Note that, if $N = 1 \wedge P_{common}(p_{a_i}, p_{v_{N+1}}) = 0$, then entropy $H = 0$. That is, if only one person is in the scenario, he is treated as the current speaker. In this situation, only temporal coincidence of proto-objects is employed for audiovisual integration.

Our previous work [50] compared the method described in Algorithm 1 with the basic approach in [49] that did not consider unseen proto-objects, in the scenario where people dynamically entered and vacated the room. When the related visual source has not yet been spotted, Algorithm 1 was able to detect the situation ($V_l = V_{N+1}$) and refused audiovisual integration, while the basic approach in [49] selected a wrong visual proto-object for audiovisual integration.

III. Online Adaptation of Audio-Motor Maps

It is essential for technical systems and living organisms to learn the mapping between audio cues and sound source location. Since this mapping is smooth for a typical robot design, it suffices to know the mapping for a couple of positions and interpolate intermediate locations. We find that a grid with a spacing of 10° in azimuth angle is well suited for practical applications. This mapping is called the audio-motor map to stress the relation to behavior. For a set of audio cues it provides the motor command such as head motion in the form of an azimuth angle to focus on the estimated position of the sound source.

In this work, the audio-motor map is represented as a simple look-up table $M_p^i(f,n)$ ($i=1$ for IID, $i=2$ for ITD) that contains the typical response of position selective filters at cue node n in frequency channel f generated by a source at true position p. Frequency channel f is in the range from $(1-100)$. Only azimuth positions p between -90° and 90° are taken into account because of the mechanical constraint of the robot head. Cue nodes n have response centers at $(-0.9,-0.8,...,0,...,0.8,0.9)$ that respond to cue values as mentioned in Section II-B. These nodes represent elements that are active for a certain range of ITD/IID values.

Fig. 3 illustrates an example of an IID map $M^1_{-90}(f,n)$ at -90°. Representing a distribution of cue values is necessary since audio cue measurements are inherently noisy due to residual background noise and echoes. Given audio-motor map $M_p^i(f,n)$ and measured cues $C'^i(f,n)$ (described in Section II-B), localization amounts to finding the position p with the highest overlap with measured cues.

The main benefit of using such a simple look-up-table, in contrast to e.g. a neural network, is that training and updating the mapping is very easy. Given the true position of the sound source and the measured distribution of cues, audio-motor maps can be updated for this position. In order to train the mapping all we need is the distribution of cues for all grid positions, that is from -90° to 90° in steps of 10° in this work. This data is normally provided in an offline calibration procedure using loudspeakers at well-known positions. For our offline calibration, we employ a fixed loudspeaker but rotate the head to different a-priori defined positions.

Unfortunately, the mapping is unlikely to be stable over extended periods of time. During the operation of a robot especially the IID part of the mapping needs to be updated continuously since it is very sensitive to small changes such as the exact position of the microphone inside its housing. For living systems the problem is even more fundamental, as they do not have a reliable calibration phase and their head is growing continuously during development.

This leads us to one of the central questions addressed in this work: how can an autonomous agent acting in a real-world environment adapt its audio-motor map? Since the measurement of audio cues is not a problem, the limiting factor is to get a reliable position estimation for the sound source, when the performance of the main mechanism via the audio-motor map is questionable.

In this work we use vision as a supporting sensory modality for the online adaptation of audio-motor maps. In contrast to previous approaches [18], [22], [23] we explicitly tackle the issue of the proper alignment of audio and visual sources through CAVSA. It is an extension of the concept of Computational Auditory Scene Analysis (CASA) [24]–[26] proposed for (robot) audition.

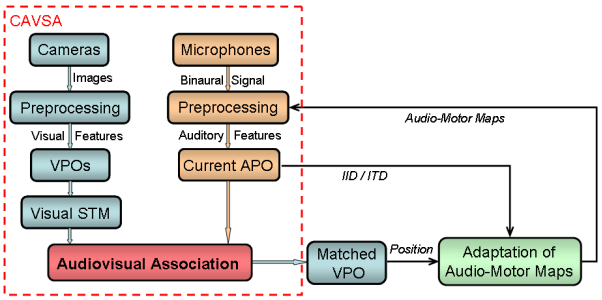

Fig. 6. Online adaptation of audio-motor maps using CAVSA. APO: audio proto-object, VPO: visual proto-object, STM: short-term memory.

A. Configuration of CAVSA

In this section we configure CAVSA for the application in online adaptation of audio-motor maps. For visual STM, the 3D world position is used to group visual proto-objects for the same speaker. Θ_s in equation 4 is set to $0.3m$, so that slight movements of speakers such as head shaking are tolerated. As shown in Fig. 6, we use the current audio proto-object together with the visual STM in the CAVSA and search for the matched visual proto-object based on Algorithm 1.

The standard deviation δ_{AV} in Eq. 7 is updated only if just one visual proto-object exists and positions near the current proto-object have been recently visually attended. The update rule is described as below:

$$\delta_{AV}^s = \begin{cases} |p_a - p_v| \cdot w + \delta_{AV}^{s-1} \cdot (1-w) & \text{if } N=1 \wedge P_u < \Theta_{P_u}, \\ \delta_{AV}^{s-1} & \text{otherwise.} \end{cases} \quad (13)$$

Here, s and w represent update step and weight, respectively. $P_u = \widehat{P}_{common}(p_a, p_{v_{N+1}})$ is the normalized probability that the current audio proto-object A is related to a visual proto-object that has not yet been seen, as described in Eq. 9. Θ_{P_u} is the threshold of P_u and is set to 0.1. δ_{AV}^0 is initialized as 40° to tolerate large position difference between audio and visual proto-object when sound localization is very inaccurate at the beginning.

B. Confidence of an Adaptation Step

The probability P_{common} (Eq. 7) for audiovisual integration is based on the position distance. Hence, the system tends to choose a wrong visual proto-object when the visual proto-object is positioned near the correct one or the performance of sound localization is poor. In both cases, P_{common} for the correct and the wrong visual proto-object tend to be similar, so that a high entropy H (Eq. 12) may refuse the audiovisual integration. However, we found in the experiments that wrong visual proto-objects in such cases can also enhance the quality of maps. This is because of the smoothness of the mapping in audio-motor maps. Thus, we define a new variable H' that

considers both entropy H and the position distance between the visual proto-objects with maximum and second maximum probability $\widehat{P}_{common}$:

$$H' = H \cdot |p_{v_1} - p_{v2}|, \tag{14}$$

where p_{v_1} and p_{v2} stand for the azimuth angles of visual proto-objects with maximal and second maximal probability, respectively. If the entropy H exceeds the threshold Θ_H, but the position distance between the visual proto-objects with maximum and second maximum probability $\widehat{P}_{common}$ is small, audio-motor maps can be updated nonetheless. In this manner the adaptation process is accelerated. Note that, when only one person is in the scenario, $p_{v_1} = p_{v2}$ and $H = 0$. The confidence c of the adaptation step is given by:

$$c = \begin{cases} 1 & \text{if } H \leqslant \Theta_H \vee H' \leqslant \Theta_{H'}, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The threshold $\Theta_{H'}$ depends on the standard deviation δ_{AV} ($\Theta_{H'} = 2 \cdot \delta_{AV}$), so that the system has a high tolerance for the visual position difference when the quality of audio-motor maps is poor.

C. Adaptation of Audio-Motor Maps

Given population-coded binaural cues $C'^i(f,n)$ in the current audio proto-object and the matched visual position $p_{v'}$, the audio-motor map $M_p^i(f,n,t)$ is updated by:

$$\begin{aligned} M_p^i(f,n,t) &= M_p^i(f,n,t-1) - F(p) \cdot \\ &\quad (M_p^i(f,n,t-1) - C'^i(f,n)), \end{aligned} \tag{16}$$

where p, i, f ,n and t stand for position, cue index, frequency channel, node and update step, respectively. The learning rate $F(p)$ is given by:

$$F(p) = c \cdot \beta \cdot \delta_{p,p_{v'}}, \tag{17}$$

where c and β represent the confidence of the matched process and the fixed adaptation rate respectively. While a high learning rate causes a more unstable learning process, convergence slows down substantially for too small learning rates. In our experiments we set $\beta = 0.2$ (based on empirical tests) to achieve a good compromise. The position evidence vector $\delta_{p,p_{v'}}$ is defined by a delta function:

$$\delta_{p,p_{v'}} = \begin{cases} 1 & \text{if } p = p_{v'}, \\ 0 & \text{if } p \neq p_{v'}. \end{cases} \tag{18}$$

IV. Results

We compared our proposed algorithm denoted as "*CAVSA*" with a heuristic method denoted as "*HEU*" which behaves similarly to methods in [22], [23] for linking auditory and visual signals. *HEU* considers the last seen face as the matched visual position to the current sound source. If more than one face appears in the camera image, the heuristic method randomly chooses one. In the scenarios we investigate there is a varying and a-priori unknown number of persons, who speak in sequence. The system initially faces the additional problem of an unreliable sound source localization due to the bad quality of the initial audio-motor maps. The system has to bootstrap itself to reach a point where sound localization can provide useful information for the audiovisual linking. Since the vision system can not cover the full environment, an utterance might be generated by a visual source outside the field of view. Additionally, hardware modifications can lead to a change in the audio-motor maps. In order to test the adaptation of our system to modifications, we modified the robot head slightly and simulated the exchange of left and right microphone. The performance of online-adapted maps is evaluated by computing the difference between online-adapted and offline-calibrated maps and also comparing sound localization results with ground truth data.

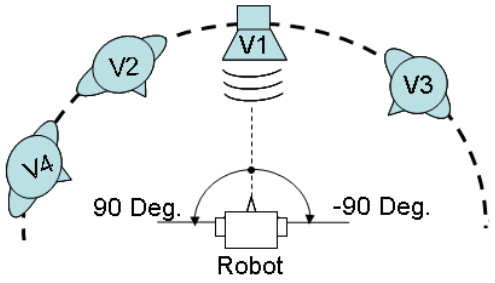

Fig. 7. Sketch of the experimental setting for bootstrapping of online adaptation. V1 is the face picture attached on the loudspeaker which is the only sound source placed at 0°, while V2, V3 and V4 are additional face pictures. **Scenario I** contains only V1; **Scenario II** contains V1 and V2; **Scenario III**: V1, V2 and V3; **Scenario IV**: V1, V2, V3 and V4

A. Performance Evaluation

The quality of an audio-motor map can be evaluated in several ways. A simple approach is to compare the map with an offline-calibrated one. The offline calibration is performed in a controlled environment where all sounds originate from a loudspeaker at a fixed position (0°). The robot head is rotated every 10° from −90° to 90°. 47 sound files of average length $4.2s$ are played at each grid position and the audio-motor maps are learned offline in a procedure outlined in [50]. The whole process, which includes sound recording, audio cue computation, map learning, takes more than two hours and requires substantial effort for the system's engineer.

Given offline-calibrated audio-motor maps, the performance of online-adapted audio-motor maps can be estimated by their normalized Euclidean distance:

$$d(M,M') = \sqrt{\frac{\sum_{p,i,f,n} \left(M_p^i(f,n) - M_p'^i(f,n)\right)^2}{K}}, \tag{19}$$

where $M_p^i(f,n)$ and $M_p'^i(f,n)$ represent online-adapted and offline-calibrated maps respectively. K is the total number of elements in an audio-motor map and satisfies $K = k_p \cdot k_i \cdot k_f \cdot k_n$, where $k_p = 19$, $k_i = 2$, $k_f = 100$ and $k_k = 19$ is the number of positions, cues, frequency channels and nodes, respectively.

As we will see later, offline-calibrated audio-motor maps are not perfect, since the mapping is not likely to be very stable. IID and ITD values vary with minor changes on the robot's hardware (e.g. microphones or head shape) so that the quality of audio-motor maps degrades. This is one of the

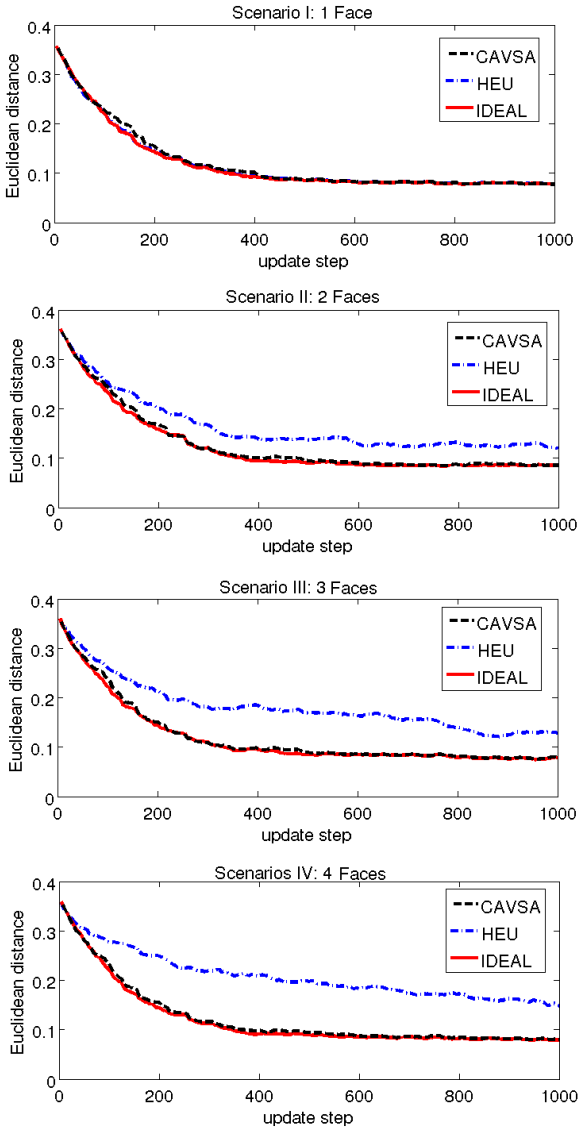

Fig. 8. Comparison of bootstrapping in different scenarios with 1-4 faces using Euclidean distance to offline-calibrated maps

main motivations for an online adaptation. In this case, offline-calibrated maps are not reliable to be used as reference, and we also apply online-adapted maps to localize sounds from a known position and measure the mean localization error. This operation can be performed in an offline test procedure or in an online process where the current sound localization result is compared with ground truth data (if available).

The mean absolute error of positions E is described by:

$$E = \frac{\sum_{t=1}^{N} |p_t - p_{true}|}{N}, \tag{20}$$

where N is the number of all measured positions. p_t represents the t-th measured position and p_{true} the ground truth position.

B. Bootstrapping of Online Adaptation

In our first experiment we started with a system that had a randomized audio-motor map. The map was filled with random numbers in the range $[-0.5, 0.5]$ using a uniform distribution.

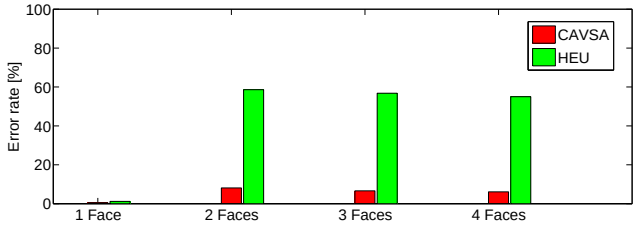

Fig. 9. Percentage of update steps where a wrong "person" is chosen as the current speaker

	1 Face	2 Faces	3 Faces	4 Faces	Offline
E	3.4°	3.9°	4.7°	4.0°	0.9°

TABLE I
Average absolute position error (E) with offline-calibrated maps and with online-adapted maps after 1000 steps in different scenarios with 1- 4 faces. Audio-motor maps were used to localize sounds in other 24 sound files that were recorded offline on each relative azimuth angle, every 10° from −90° to 90°.

Therefore, it was not able to reliably map audio cues to source positions. Since there were also no other cues that could guide the system to which visual proto-object corresponds to the sound source, improving the audio motor map was difficult. We made a set of experiments to analyze this situation and found that our system was capable of bootstrapping. Instead of a real person in the experiments, we used a loudspeaker on which a picture of a face was attached, as the experiments were very time-consuming. The loudspeaker was placed in front of the robot (0°), at a distance of $1m$ and at the same height, as shown in Fig. 7. During online adaptation, the robot head oriented itself to a random horizontal angle in the range $[-90°, 90°]$ after finishing each adaptation step. The random head motion was used to simulate the head motion during the normal operation of the robot. The acquisition of auditory and visual signals was interrupted during head movement, so that audio-motor maps were only adapted in still status. Up to 3 additional pictures of a face (V2, V3 and V4 in Fig. 7) were attached on objects that were also placed at a distance of $1m$ and at the same height as the robot head. Although the faces are frontal to the robot, the robot can not see all the faces at the same time due to limited camera field of view and head movements. Since the only sound source was the loudspeaker, the ground truth position was 0°. For these scenarios, our system *CAVSA* was compared with the heuristic method *HEU* which is similar to methods in [22], [23] for linking auditory and visual signals. Also the method using known sound source position (0°) is shown as reference and denoted as "*IDEAL*".

Fig. 8 compares the bootstrapping in four different scenarios with 1-4 faces using Euclidean distance to offline-calibrated maps. It took about 72 minutes for every 1000 adaptation steps. The percentage of update steps where a wrong "person" was chosen as the current speaker is shown in Fig. 9. We can see that *CAVSA* was able to bootstrap itself in different scenarios and nearly reached the performance of *IDEAL*. *CAVSA* overperformed *HEU* particularly when more than one face was in the scenario, because *HEU* tended to select a wrong visual proto-object as shown in Fig. 9. In

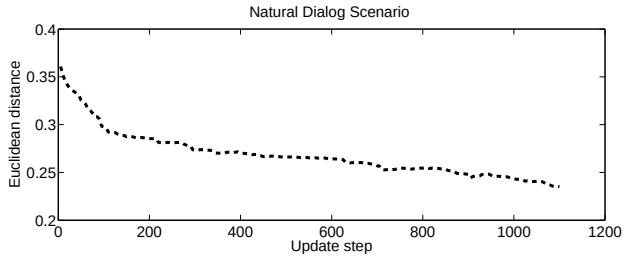

Fig. 10. A natural dialog scenario where four persons spoke freely without any script: Euclidean distance between online-adapted and offline-calibrated maps

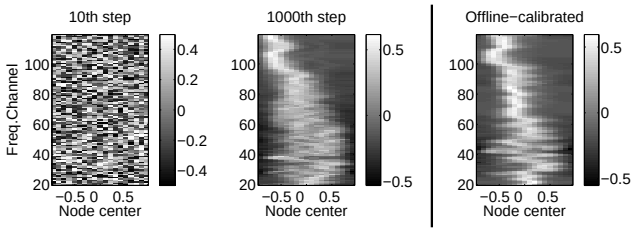

Fig. 11. In the natural dialog scenario with four speakers: online-adapted IID maps of -20° after 10 and 1000 steps. Offline-calibrated maps are used as reference.

comparison, *CAVSA* had only a small error rate mainly due to inaccurate sound localization. The small error rate of *CAVSA* in the scenario with one face was because of falsely detected faces. There should be no selection error in the one-person scenario, because only temporal coincidence of proto-objects is employed for audiovisual integration in this case and the person is treated as the current sound source. As shown in Fig. 8, the convergence of Euclidian distance is nearly achieved after 400 steps, which took about 30 minutes, while offline calibration requires more than 2 hours.

We also verified the quality of online-adapted maps using them to localize sounds from a known position. The average absolute position error after 1000 steps in different scenarios are shown in TABLE I. The performance of offline maps is also given as reference. We can see that the results were consistent using different performance metrics.

Moreover, we tested the learning of the audio-motor maps from a random initialization using only human speakers for 50 minutes. Four participants entered the room and spoke freely without any script. They faced the robot most of the time, spoke sometimes concurrently and were free to move. This scenario is very challenging for our system, because sound source separation is missing in the current system. When people talk at the same time, the sound position is measured as some average of all active speakers and the uncertainty of audiovisual integration can be very high. In such confusing situations, entropy H can exceed its threshold Θ_H and audiovisual integration is refused. Fig. 10 shows the Euclidean distance between online-adapted and offline-calibrated maps over time. We found that the system was capable of bootstrapping even in this scenario and the adaptation progress was steady. The average absolute position errors were measured in the same manner as that in TABLE I.

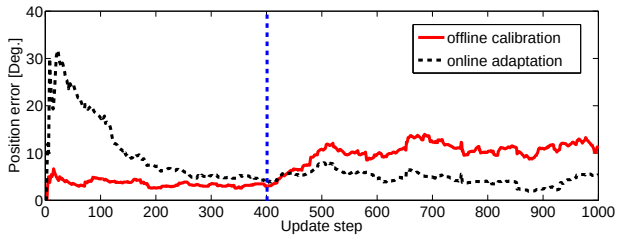

Fig. 12. Modification of robot head in the 400th step: Mean position error with offline-calibrated and online-adapted maps in **scenario I**

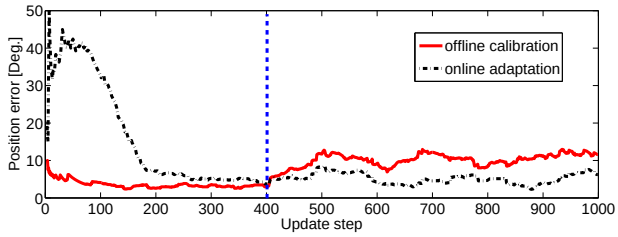

Fig. 13. Modification of robot head in the 400th step: Mean position error with offline-calibrated and online-adapted maps in **scenario II**

The errors after 10 and 1000 update steps are 61° and 22° , respectively. Fig. 11 shows for example the online-adapted IID maps of -20° after 10 and 1000 steps, as well as the offline-calibrated IID map of -20°. The structure of the online-adapted IID map after 1000 steps is similar as the offline-calibrated map. In comparison to scenarios as shown in Fig. 7, the learning progress was slow due to many complex and confusing situations. Therefore the maps need to be adapted longer than in relatively simple scenarios.

C. Adaptation Ability

In another set of experiments, we analyzed how the system responded to deviations in the robot's hardware that induce a change in the audio-motor map. Smaller modifications are quite common, especially in larger projects where the audio engineer is not the only person to work with the robot. The audio-motor maps were adapted online in the standard hardware configuration as in **scenario I** with only one face or in **scenario II** with two faces (Fig. 7). In the 400th step, we applied or simulated hardware modifications that induced a small but measurable change in the audio-motor map. Then the audio-motor maps were adapted continuously in the same scenario. Since sounds were generated from a defined speaker position (0°), we were able to measure sound localization performance using the current online-adapted audio-motor maps and the offline-calibrated. For each update step, we estimated the mean position error in degrees using Eq. 20 for the last 100 steps. The position errors are averaged because the localization performances of an audio-motor map are very different for each position and are normally better for center positions and a single measurement can not represent the performance of the current audio-motor map.

1) Modification of robot head: We modified the robot head slightly by putting a paper hat over the outer ears to alter the mapping between audio cues and sound position. Fig.

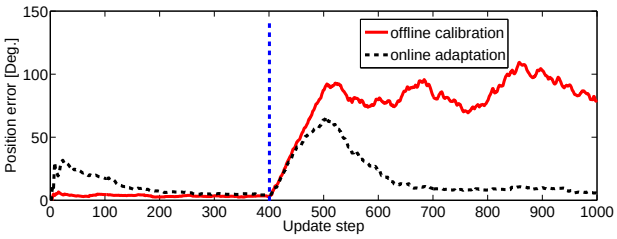

Fig. 14. Simulated exchange of two microphones in the 400th step: Mean position error with offline-calibrated and online-adapted maps in **scenario I**

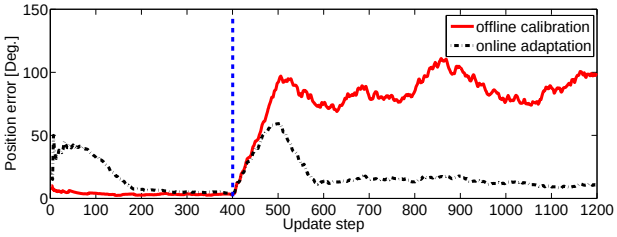

Fig. 15. Simulated exchange of two microphones in the 400th step: Mean position error with offline-calibrated and online-adapted maps in **scenario II**

12 shows the mean position error with offline-calibrated and online-adapted maps in scenario I with only face. We notice that the error increase for both maps was not sudden but rather slow after the 400th step due to the averaging of position errors over time. Then performance of online-adapted maps recovered over time, while the mean position error of offline-calibrated maps was always about 10°. Fig. 13 shows the comparison of sound localization performance between offline-calibrated and online-adapted maps in scenario II with two faces. In comparison to Fig. 12, we can see that the performance difference of online-adapted maps between two scenarios is not significant. In the multi-person scenario, the percentage of update steps using a wrong visual proto-object was only 7%.

2) Exchange of two microphones: We simulated the exchange of left and right microphone in the 400th update step to observe if the mapping is able to adapt to the new situation. With this extreme modification, the sound localization system detects a sound from the right side as coming from the left side. Fig. 14 (resp. Fig. 15) illustrates the sound localization performance of offline-calibrated and online-adapted maps in scenario I with a single face (resp. in scenario II with two faces). Both online-adapted and offline-calibrated maps deteriorated after exchanging microphones. Then online-adapted maps recovered quickly while the sound localization results using offline-calibrated maps were completely wrong. In scenario I, the mean position error of online-adapted maps after 1000 steps was 5.8°. That is, the performance reverted already to its old localization performance. In comparison, the mean position error of online maps in the multi-person scenario was 13.5° after 1000 steps and 10.4° after 1200 steps because of many wrongly selected visual proto-objects. In this scenario, the percentage of update steps where a wrong visual proto-object was chosen was 38%. This was not surprising as the modification was extreme, and the maps need to be adapted continuously to achieve its old performance.

In the above experiments, we observed that the localization performance of the online-adapted maps closely approached the localization performance of the offline-calibrated one before hardware modifications. After modifications the performance for both maps decreased. However, the performance of online-adapted audio-motor maps increased over time. Without online adaptation we would have to repeat the whole process of offline calibration that takes 2 hours. This is one main motivation to online adapt audio motor-maps.

V. Discussion

Inspired by the important role of vision for the calibration of audio-motor maps in animals and humans, several recent works [22], [23] use visual feedback to provide the sound source position. However, they employ heuristics for linking visual and auditory information and thus can only work in limited environments. In contrast, we applied CAVSA to search for the matched visual source to the current sound in multi-person environments. In CAVSA, the scene is represented with auditory and visual features in form of proto-objects. Proto-objects for the same speaker are grouped together in auditory and visual STMs using position information. Finally, audiovisual integration is based on either temporal or temporal and spatial coincidence of auditory and visual stimuli.

Comparing our approach with the state of the art in terms of learning progress, we have shown that our approach with CAVSA was more robust in multi-person scenarios. We have also shown that our system was able to bootstrap with a randomized audio-motor map in multi-person environments. The number of persons (2-4) has little influence on the learning performance. The percentages of update steps where a wrong person is chosen as the current speaker were about 6% in scenarios with 2-4 persons. Moreover, our system was capable of bootstrapping itself in a natural dialog scenario with 4 human speakers. The learning progress was slow but steady in this complex scenario. In case of hardware modifications, the online-adapted maps were able to revert to their old localization performance in the single-person scenario. In the multi-person scenario, online-adapted maps nearly reached the same performance, when the mapping was slightly modified by putting a hat on the robot head. However, when the mapping was changed so drastically that the left and right side were confused as in the experiment with the simulated exchange of the two microphones, the system tended to choose a wrong visual proto-object due to inaccurate sound localization especially shortly after the hardware modification. The recovery period was longer than for minor modifications of the mapping. But this modification is rather extreme.

When the sound localization is inaccurate or persons stand near the current speaker, the system tends to choose a wrong visual proto-object. We found that wrong visual proto-objects in these cases can also enhance the quality of audio-motor maps because of the smoothness of the mapping. However, wrong visual proto-objects may have negative impacts in other applications. In a dialog system, it is confusing for the dialog manager when words are assigned to wrong persons. In the real

world, animals may not be able to avoid danger using wrong audiovisual integration. Therefore, we defined variable H' and confidence c in Section III-B to apply incorrect but useful visual proto-objects for adaptation of audio-motor maps, while Θ_H, the threshold of the uncertainty of audiovisual integration, could be set dependent on other tasks such as assignment of words to corresponding speakers. In multi-person scenarios, the larger the threshold Θ_H ($\Theta_H \in [0,1]$) and $\Theta_{H'}$ are, the more possible matched but also the more wrong visual proto-objects might be used for online adaptation. Hence the tradeoff between quality and speed of the online adaptation is faced. We choose thresholds dependent on situations. In case of a service robot that interacts often with people in the daily life, the system has enough time and samples for online adaptation. For accuracy, we could adapt audio-motor maps when only one person is in the scenario. That is, we could set both Θ_H and $\Theta_{H'}$ in Eq. 15 to 0, so that online adaptation is stopped in multi-person scenarios. In contrast, in our experiments with less than 2 hours, we do not have enough samples and thus make a compromise to use also some wrong visual proto-objects that can enhance the quality of audio-motor maps. This is obviously better than just waiting for an one-person scenario, although the performance increase using the wrong visual proto-object is slower than using the correct one. In this work, Θ_H is set to 0.8 based on empirical test, and $\Theta_{H'}$ adapts to the standard deviation δ_{AV} ($\Theta_{H'} = 2 \cdot \delta_{AV}$, δ_{AV} is defined in Eq. 13) so that the system has a high tolerance for the visual position difference when the quality of audio-motor maps is poor. Furthermore, the use of uncertainty information (entropy H) is application dependent. If the visual proto-object V_l defined in Eq. 10 is unreliable for audiovisual integration or probably has not yet been seen, our system does not adapt maps in this step. However, in a dialog system the decision has to be made who is currently speaking. The uncertainty information (entropy H) could be sent to the dialog manager. If an audiovisual association is unreliable, the dialog manager may ask for confirmation.

Compared to offline calibration, the main benefit of online adaptation is that the maps can be quickly adapted to the modifications of acoustic conditions. For example in the experiment where a paper hat was put on top of the robot, the recovery period of online maps was about 15 minutes (from the 400th to the 600th step as shown in Fig. 12 and Fig. 13), while 2 hours have to be spent for repeating offline calibration in this case. The temporal length of adaptation steps depends mainly on the length of the sound segments and how often the robot hears a sound, because the run time of an adaptation step is less than $1s$ and is normally less than a sound segment so that the system has to wait for the next audio proto-object after the computation is finished. For example, it took 72 minutes for 1000 adaptation steps using the recorded sound files of average length $4.2s$ (Fig. 8), while only 50 minutes were spent for 1100 steps in a natural dialog scenario (Fig. 10), because the participants talked all the time in the experiment and generated audio proto-objects faster.

Our approach does not need control over the robot motion and thus can work during other tasks. However, motion has an important influence on the online adaptation of audio-motor maps. First, motion of the robot head or persons is necessary for the online adaptation of audio-motor maps. Audio-motor maps should be updated for all azimuth angles (sound positions relative to the robot head). We assume that most relative azimuth angles are generated through the robot's head motion and movements of the interacting partners during various tasks. In our experiments, the random head motion was used to simulate the robot's head motion during the normal operation of the robot. However, given the extreme situation where a robot and a person do not move, but just look at each other and talk all the time, the robot could update maps only for some center azimuth angles. Of course, we can not expect the robot to localize a sound from an azimuth angle, from where it has never heard a sound and learned the audio-motor map before. Moreover, certain motions support updating the scene representation and thus indirectly enhance the performance of online-adapted maps [66]. For example our system is capable of detecting the situation that the current sound is related to a person who has not yet been seen. For this case, audiovisual integration is refused. If the robot is required to turn to a sound in certain tasks, the visual proto-object of the unseen speaker will be stored in the STM and the scene representation is updated.

There is a number of issues we do not tackle in this work and which need to be addressed in future work. We assume that all sounds originate from human speakers who orient their faces to the robot. Environmental sounds such as sounds generated by cars could not be handled properly because the system can not visually extract cars and it does not know which type of visual proto-object corresponds to which sounds. The latter problem could be solved when the system learns the typical sound features for different types of visual objects (i.e. learning the correlation between audio and visual features in the proto-objects). As a short-cut, our system employed filters for audio proto-objects based on very simple cues that remove most environmental sounds [62].

In future work, we will test online adaptation of audio-motor maps in the system [67] that aims at assigning words to corresponding speakers using CAVSA in natural dialog scenarios with multiple people. For the different task of analyzing "who is currently speaking", [67] used more visual features in CAVSA such as color and texture features of different upper body parts (clothes, collar and hair) to recognize persons again, when they leave and enter the scene, or move outside the camera field of view. Additionally, we plan to use more auditory features in CAVSA to recognize voices, and expect that once an audio and a visual proto-object are integrated, the integration can be maintained for a certain time. The stabilized audiovisual integration in CAVSA would enhance the performance of online adaptation of audio-motor maps.

ACKNOWLEDGMENT

The authors would like to thank Mark Dunn, Bram Bolder, and Stephan Kirstein for their helpful support with the vision system and all the colleagues at Honda Research Institute Europe, who participated in the experiments. This work was supported by the Honda Research Institute.

REFERENCES

[1] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using binaural and spectral cues for azimuth and elevation localization," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2008.

[2] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, 3rd ed. The MIT press, USA-Cambridge MA, 2001.

[3] K. A. Davis, R. Ramachandran, and B. J. May, "Auditory processing of spectral cues for sound localization in the inferior colliculus," *JARO*, vol. 04, pp. 148–163, 2003.

[4] D. Gill, L. Troyansky, and I. Nelken, "Auditory localization using direction-dependent spectral information," *Neurocomputing*, vol. 32–33, pp. 767–773, 2000.

[5] M. M. V. Wanrooij and A. J. V. Opstal, "Contribution of head shadow and pinna cues to chronic monaural sound localization," *Journal of Neuroscience*, vol. 24, no. 17, pp. 4163–4171, 2004.

[6] T. Rodemann, "Spectral cues to source position in robots with arbitrary ear shapes," in *Proc. Int. Conf. on Advanced Robotics (ICAR)*, 2011.

[7] H. Viste and G. Evangelista, "Binaural sound localization," *In Proc. Int. Conf. on Digital Audio Effects (DAFx'04)*, 2004.

[8] T. Rodemann, "A study on distance estimation in binaural sound localization," in *Proc. of the IEEE/RSJ conference on Intelligent Robots and Systems (IROS)*, 2010.

[9] J. Weng and K. Y. Guentchev, "Three-dimensional sound localization from a compact non-coplanar array of microphones using tree-based learning," *Journal of the Acoustical Society of America*, vol. 110, pp. 310–323, 2001.

[10] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner, "A probabilistic model for binaural sound localization," *IEEE Transactions on Systems, Man and Cybernetics*, 2006.

[11] E. I. Knudsen and P. F. Knudsen, "Vision calibrated sound localization in developing barn owls," *The Journal of Neuroscience*, vol. 9, pp. 3306–3313, 1989.

[12] A. J. King, M. E. Hutchings, D. R. Moore, and C. Blakemore, "Developmental plasticity in the visual and auditory representation in the mammalian superior colliculus," *Nature*, vol. 332, pp. 73–76, 1988.

[13] P. S. Hyde and E. I. Knudsen, "The optic tectum controls visually guided adaptive plasticity in the owl's auditory space map," *Nature*, vol. 415, pp. 73–76, 2002.

[14] P. M. Hofman, J. G. A. V. Riswick, and A. J. van Opstal, "Relearning sound localization with new ears," *Nature neuroscience*, vol. 1, pp. 417–421, 1998.

[15] J. I. Gold and E. I. Knudsen, "Abnormal auditory experience induces frequency-specific adjustments in unit tuning for binaural localization cues in optic tectum of juvenile owls," *Journal Neuroscience*, vol. 20, no. 2, pp. 862–877, 2000.

[16] O. Kacelnik, F. Nodal, C. Parsons, and A. King, "Training-induced plasticity of auditory localization in adult mammals," *Public Library of Science (PLoS) Biology*, vol. 4, no. 4, pp. 627–638, 2006.

[17] W. M. DeBello and E. I. Knudsen, "Multiple sites of adaptive plasticity in the owl's auditory localization pathway," *The Journal of Neuroscience*, vol. 24, no. 31, pp. 6853–6861, 2004.

[18] T. Rodemann, K. Karova, F. Joublin, and C. Goerick, "Purely auditory online-adaptation of auditory-motor maps," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2007.

[19] M. P. Zwiers, A. J. V. Opstal, and J. R. M. Cruysberg, "A spatial hearing deficit in early-blind humans," *Neuroscience*, vol. 21, pp. 1–5, 2001.

[20] E. I. Knudsen, "Instructed learning in the auditory localization pathway of the barn owl," *Nature*, vol. 417, pp. 322–328, 2002.

[21] ——, "Early blindness results in a degraded auditory map of space in the optic tectum of the barn owl," *Proceedings of the National Academy of Sciences, USA*, vol. 85, pp. 6211–6214, 1998.

[22] J. Hörnstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots-building audio-motor maps based on the HRTF," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2006.

[23] H. Nakashima and T. Mukai, "3D sound source localization system based on learning of binaural hearing," *IEEE Int. Conf. on Systems, Man and Cybernetics*, 2005.

[24] D. F. Rosenthal and H. G. Okuno, Eds., *Computational Auditory Scene Analysis*. Lawrence Erlbaum Assoc Inc, 1998.

[25] P. Divenyi, Ed., *Speech separation by humans and machines*. Kluwer Academic Publishers, 2004.

[26] D. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press/Wiley-Interscience, 2006.

[27] J. M. Schmüdderich, "Multimodal learning of grounded concepts in embodied systems," Dissertation, Bielefeld University, 2010.

[28] B. Bolder, M. Dunn, M. Gienger, H. Janssen, H. Sugiura, and C. Goerick, "Visually guided whole body interaction," *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2007.

[29] T. Rodemann, F. Joublin, and C. Goerick, "Audio proto objects for improved sound localization," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2009.

[30] T. H. Weisswange, C. A. Rothkopf, T. Rodemann, and J. Triesch, "Bayesian cue integration as a developmental outcome of reward mediated learning," *PLoS ONE*, vol. 6, p. 11, 2011.

[31] T. Noesselt, J. W. Rieger, M. A. Schoenfeld, M. Kanowski, H. J. Heinze, and J. Driver, "Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices," *Neuroscience*, vol. 27, pp. 11 431–11 441, 2007.

[32] D. Alais and D. Burr, "The ventriloquist effect results from near-optimal bimodal integration," *Current Biology*, vol. 14, pp. 257–262, 2004.

[33] M. T. Wallace, G. E. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan, and J. A. Schirillo, "Unifying multisensory signals across time and space," *Experimental Brain Research*, vol. 158, pp. 252–258, 2004.

[34] E. I. Knudsen and P. K. Knudsen, "Vision guides the adjustment of auditory localization in young barn owls," *Science*, vol. 230, pp. 545–548, 1985.

[35] M. T. Wallace and B. E. Stein, "Early experience determines how the senses will interact," *J Neurophysiol*, vol. 97:921-926, 2006.

[36] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," *Advances in Neural Information Processing Systems*, vol. 12, pp. 813–819, 2000.

[37] H. Hung and G. Friedl, "Towards audio-visual on-line diarization of participants in group meetings," *In Proc. of European Conference on Computer Vision (ECCV)*, 2008.

[38] A. Noulas and B. J. A. Krose, "On-line multi-modal speaker diarization," *In Proc. Int. Conf. on Multimodal interfaces (ICMI)*, 2007.

[39] A. L. Casanovas, G. Monaci, and P. Vandergheynst, "Blind audiovisual source separation using sparse representations," *IEEE Transactions on Multimedia*, pp. 358 – 371, 2010.

[40] G. Friedland, C. Yeo, and H. Hung, "Visual speaker localization aided by acoustic models," *ACM Multimedia*, pp. 195–202, 2009.

[41] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 6, pp. 406–413, 2004.

[42] H. Kim, K. Komatani, T. Ogata, and G. Okuno, "Auditory and visual integration based localization and tracking of humans in daily-life environments," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2007.

[43] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for humanoids," *Proc. of 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

[44] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell, "A probabilistic framework for multi-modal multi-person tracking," *Conference on Computer Vision and Pattern Recognition Workshop*, 2003.

[45] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud, "Audio-visual clustering for multiple speaker localization," *5th International Workshop on Machine Learning for Multimodal Interaction*, 2008.

[46] K. Bernardin and R. Stiefelhagen, "Audio-visual multi-person tracking and identification for smart environments," *Int. Proc. of the 15th Int. Conf. on Multimedia*, 2007.

[47] M. Aytekin, C. F. Moss, and J. Z. Simon, "A sensorimotor approach to sound localization," *Neural Computation*, vol. 20, pp. 603–635, 2008.

[48] J. F. Bergan, P. Ro, D. Ro, and E. I. Knudsen, "Hunting increases adaptive auditory map plasticity in adult barn owls," *The Journal of Neuroscience*, vol. 25, no. 42, pp. 9816–9820, 2005.

[49] R. Yan, T. Rodemann, and B. Wrede, "Computational audiovisual scene analysis for dialog scenarios," *IROS 2011 Workshop on Cognitive Neuroscience Robotics, http://www.honda-ri.de/intern/Publications/PUBA-22*, 2011.

[50] ——, "Learning of audiovisual integration," *In Proc. of IEEE Int. Conf. on Development and Learning (ICDL)*, 2011.

[51] H. G. Okuno, K. Nakadai, T. Lourens, and H. Kitano, "Sound and visual tracking for humanoid robot." *Appl. Intell.*, vol. 20, no. 3, pp. 253–266, 2004.

[52] D. Burr and D. Alais, "Combining visual and auditory information," *Progress in Brain Research*, 2006.

[53] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[54] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmüdderich, and C. Goerick, "Expectation-driven autonomous learning and interaction system," in *In Proc. of IEEE-RAS Int. Conf. on Humanoid Robots*, 2008.
[55] T. Stein, M. V. Peelen, and P. Sterzer, "Adults' awareness of faces follows newborns' looking preferences," *PLoS ONE*, vol. 6, p. 1, 2011.
[56] T. Farroni, M. H. Johnson, E. Menon, L. Zulian, D. Faraguna, and G. Csibra, "Newborns' preference for face-relevant stimuli: Effects of contrast polarity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 16 907–16 908, 2005.
[57] J. Ramírez, J. M. Górriz, and J. C. Segura, "Voice activity detection. Fundamentals and speech recognition system robustness," *System*, pp. 1–22, 2007.
[58] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filterbank," *Apple Computer Co., Technical Report*, vol. 35, 1993.
[59] M. Heckmann, T. Rodemann, F. Joublin, and C. Goerick, "Auditory inspired binaural robust sound source localization in echoic and noisy environments," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2006.
[60] T. Rodemann, M. Heckmann, B. Schoelling, F. Joublin, and C. Goerick, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2006.
[61] A. Pouget, P. Dayan, and R. Zemel, "Information processing with population codes," *Nature Neuroscience Reviews*, vol. 1, pp. 125–132, 2000.
[62] T. Rodemann, F. Joublin, and C. Goerick, "Filtering environmental sounds using basic audio cues in robot audition," *In Proc. Int. Conf. on Advanced Robotics (ICAR)*, 2009.
[63] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.
[64] T. Rodemann, F. Joublin, and C. Goerick, "Continuous and robust saccade adaptation in a real-world environment," *KI-Künstliche Intelligenz*, 2006.
[65] K. M. Visscher, E. Kaplan, M. J. Kahana, and R. Sekuler, "Auditory short-term memory behaves like visual short-term memory," *PLoS Biology*, vol. 5, no. 3, 2007.
[66] C. Karaoguz, T. Rodemann, B. Wrede, and C. Goerick, "Learning information acquisition for multi-tasking scenarios in dynamic environments," *IEEE Transactions on Autonomous Mental Development*, vol. 5, pp. 46–61, 2013.
[67] R. Yan, T. Rodemann, and B. Wrede, "Simple auditory and visual features for human-robot dialog scene analysis," *In Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012.



Rujiao Yan received the Diploma degree in Electronical Engineering from the Technical University Dortmund in 2008. Since 2010, she has been working toward the Ph.D. degree at the Research Institute for Cognition and Robotics (CoR-Lab) in Bielefeld University and is a guest scientist at the Honda Research Institute Europe (HRI-EU). Her research interests are human-robot interaction, multimodal perception and biologically inspired robots.



Tobias Rodemann studied physics and neuroinformatics at the Ruhr Universität Bochum, Germany, and received his Dipl.-Phys. degree from the Universität Bochum in 1998 and a Ph.D. degree from the Technische Universität Bielefeld, Germany in 2003. Since 1998 he is working at the Honda Research Institute Europe, Offenbach, Germany. Previous research fields were evolutionary algorithms, biologically-inspired vision systems, information processing with spiking neurons and learning of sensory-motor maps. Since 2003 he is working as a senior scientist on sound localization, auditory scene analysis and audiovisual interaction.



Britta Wrede is head of the research group Hybrid Society within the Institute for Cognition and Robotics (CoR-Lab) at Bielefeld University. She received her Masters degree in Computational Linguistics and the Ph.D. degree (Dr.-Ing.) in computer science from Bielefeld University in 1999 and 2002, respectivley. From 2002 till 2003 she pursued a Post-Doc program of the DAAD at the speech group of the International Computer Science Institute (ICSI) in Berkeley, USA. In 2003 she rejoined the Applied Informatics Group at Bielefeld University and was involved in several EU and national (DFG, BMBF) projects. Since 2008 she is heading her own research group at the CoR-Lab. Britta Wrede is interested in speech recognition, prosodic and acoustic speech analysis for propositional and affective processing, and dialog modeling as well as human-robot interaction. Her current research focuses on the integration of multimodal information as a basis to bootstrap speech and action learning in a tutoring scenario.