COMPUTATIONAL AUDIOVISUAL SCENE ANALYSIS

by

Rujiao Yan

Dissertation

submitted to the
Faculty of Technology at Bielefeld University
for the degree of

Doktor der Ingenieurwissenschaften
(Dr.-Ing.)

A dissertation submitted to the Faculty of Technology at Bielefeld University for the degree of Doktor - Ingenieurwissenschaften (Dr.-Ing.) on May 01, 2014.

Reviewed by

- Prof. Dr. Britta Wrede

  Bielefeld University

- Dr. Tobias Rodemann

  Honda Research Institute Europe

- Asst. Prof. Dr. Gokhan Ince

  Istanbul Technical University

Accepted on August 29, 2014, on behalf of the Faculty of Technology at Bielefeld University, Germany, by the following dissertation committee:

Prof. Dr. Barbara Hammer	Bielefeld University
Prof. Dr. Britta Wrede	Bielefeld University
Dr. Tobias Rodemann	Honda Research Institute Europe
Asst. Prof. Dr. Gokhan Ince	Istanbul Technical University
Dr. Christina Unger	Bielefeld University

Acknowledgments

First, I would like to gratefully and sincerely thank my supervisor, Tobias Rodemann, for his guidance, support, understanding and patience during my doctoral studies. I am truly thankful for his selfless dedication to both my personal and academic development. I cannot think of a better supervisor to have.

I also want to express my sincere thanks to my supervisor, Prof. Britta Wrede, for her continuous valuable support and guidance in the last years despite my geographical remoteness.

I would also like to thank Mark Dunn, Bram Bolder and Stephan Kirstein for their helpful support with the vision system and all the colleagues at the Honda Research Institute Europe (HRI-EU) who participated in the experiments.

I really appreciated the interesting and useful discussions with these people, and also with Sven Rebhan, Thomas Weisswange, Ursula Körner, Heiko Wersing, Martin Heckmann, Frank Joublin, Claudius Gläser, Irene Ayllón Clemente and Samuel Kevin Ngouoko.

I also wish to thank Stephan Kirstein, Thomas Weisswange and Bram Bolder for their willingness, time and effort to review my thesis, and their valuable feedback on an earlier version of this thesis.

Furthermore, the work presented in this thesis was a cooperation project between the Research Institute for Cognition and Robotics (CoR-Lab) at the Bielefeld University and HRI-EU in Offenbach/Main. I would like to thank the heads of both institutes Prof. Jochen Steil and Prof. Bernhard Sendhoff for giving me the opportunity to pursue my interest in robotics in an excellent environment.

Last but definitely not the least, I would like to acknowledge the endless support, encouragement, tolerance and understanding of my husband, my mother and brother. Without them, I do not think that I could overcome the difficulties during these years.

Abstract

In the field of robot audition the topic of Computational Auditory Scene Analysis (CASA) has been studied since many years. However, the focus of most CASA research 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 Computational AudioVisual Scene Analysis (CAVSA) proposed in this work 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 dialogue scenario with an unknown number of speakers who have not been seen or heard before. In addition to the grouping issue, it is also estimated how many interaction partners are in the scenario and where they are. Moreover, CASA is extended to CAVSA by combining visual and auditory modalities for achieving a more accurate and reliable perception. CAVSA can be used as a complementary analysis step for applications such as speech recognition.

CAVSA is a challenging task due to the complexity of dialogue scenarios. First, speakers are unknown in advance, thus a database for training high-level features beforehand to recognize faces or voices is not available. Second, people can dynamically come into and leave the scene, may move all the time and even change their locations outside the camera field of view. Third, the robot can not see all the people at the same time due to limited camera field of view and head movements. Moreover, a sound could be related to a person who stands outside the camera field of view and has never been seen.

The main contribution of this thesis lies in the system architecture. Three important concepts are applied together in the system: proto-object, Short Term Memory (STM) and audiovisual association. In this work, proto-objects are considered as a compressed form of various features. They can basically combine an arbitrary number of features and can be easily extended with new features to improve performance without changing the architecture of the system. STM can represent environment changes such as changes in number and position of speakers, and acquire knowledge of new speakers continually in an unsupervised manner. Finally, the current auditory signal is linked to the matched visual signal to answer the question “Who is speaking?”. The second contribution is combining many simple auditory and visual features to represent speakers, since training high-level features beforehand to recognize faces or voices is impossible. Although each individual feature is normally not good enough to differentiate between speakers, their combination can fulfill this task.

This work applies CAVSA in two applications. First, in online adaptation of audio-motor maps, CAVSA is applied for finding the matched visual source to the current audio source in multi-person environments, thus vision can be used to provide precise position information. Only a small set of simple features is used in the original version of CAVSA, since it is not necessary to identify persons in this application. It is shown that the approach with CAVSA is more robust in multi-person scenarios than the state of the art in terms of learning progress. Then CAVSA involving more complicated issues is used in human-robot dialogue scenarios. In this application, the system needs to integrate more features to recognize speakers. It is shown that the CAVSA system can assign words to corresponding speakers when multiple people dynamically enter and leave the room. In comparison to the state of the art, a person can be recognized again even when disappearing for a while and then reappearing. Moreover, typical environmental sounds in offices and households can be filtered out based on simple auditory features.

The system used only a humanoid robot head with a pair of cameras and a pair of microphones. The head is mounted on a pan-tilt unit.

Contents

1. Introduction

Robots have been available since a long time to perform repetitive and dangerous tasks in industrial settings. With the rapid development of robot research, service robots for personal and domestic use are playing a more and more important role. According to the International Federation of Robotics, about 2.5 million service robots for personal and domestic use were sold worldwide in 2011 [IFR, 2012]. It is further projected that the sales of all types of personal service robotics could reach about 15.6 million units in the period 2012-2015. These include robots for floor cleaning, lawn-mowing, entertainment, education, assistance for handicapped people, etc. Will robots become part of our day-to-day lives in the future? Bill Gates gave his prediction related to this question [Gates, 2007]. He linked the current status of robotic technology to the computer industry during the mid-1970s, around the time that he and his fellow entrepreneur Paul Allen launched Microsoft. He also predicted that the next hot field will be robotics and wrote: "I can envision a future in which robotic devices will become a nearly ubiquitous part of our day-to-day lives." [Gates, 2007]

Although it is difficult to say how soon this day will come, there is a clear tendency for robots to be more of a companion to people and no longer just a tool. In the future, they will probably help people with all the household chores, provide physical assistance and even companionship for the elderly. All the aspects require friendly Human Robot Interactions (HRI). One goal in HRI is natural interaction which means that people communicate with robots in a way similar to human-human interaction. This makes robots more natural and acceptable for humans. Humans communicate with other people using verbal signals and nonverbal signals (e.g. body gesture, gaze and facial expression). In order to emulate human communication capabilities, robots must be able to process and produce such signals. There has been a tremendous interest in these issues. Among them, the capability of spoken language processing is a very important issue in the natural HRI, since robots should at least allow humans to tell them what to do.

There are a lot of factors influencing the performance of spoken language communication. Background noise is one factor investigated by many researchers. Even the best speech

recognition system requires that the speech is produced in a fairly quiet background. This is because the system can not distinguish between the sounds and would consider the mixture of the signal with its background as a single sound. To overcome this issue, there are two very different fields with opposing viewpoints related to sound source separation. One is the field of multichannel statistical methods such as Blind Source Separation (BSS), which usually use more than two microphones, see for instance [Cichocki and Amari, 2002]. The other is the field of psychoacoustic-based scene analysis. Humans can easily deal with the "cocktail party problem", following a single speaker's words while a number of people are talking simultaneously in a room with loud background music. Bregman [Bregman, 1990] argues that the sound reaching our ears is processed by Auditory Scene Analysis (ASA). ASA is a psychoacoustic concept and aims to separate auditory streams from their mixture, such that each stream represents a single sound source in the acoustic environment. In the field of robot audition the topic of building an ASA has been studied since many years. Recent books on Computational Auditory Scene Analysis (CASA) describe intensive work in this field [Rosenthal and Okuno, 1998, Divenyi, 2004, Wang and Brown, 2006]. CASA is the field of computational study that aims to achieve human performance in ASA by using only one or two microphone recordings of the acoustic scene.

Another main factor influencing the performance of language communication is the grouping of the individual sounds according to their respective sources (esp. grouping different utterances from different speakers). Even if all people speak in sequence and speech recognition performs perfectly, it is desirable to assign words to matched speakers. Otherwise, the robot could misinterpret the dialogue. However, most current work of CASA focuses on sound source separation and this issue is often neglected. Therefore, the work proposed in this thesis focuses on the issue of grouping. Moreover, CASA is extended to Computational AudioVisual Scene Analysis (CAVSA) by combining visual and auditory modalities for achieving a more accurate and reliable perception. CAVSA is a very challenging task and analyzes a dialogue scenario with an unknown number of speakers who have not been seen or heard before. In addition to the grouping issue, it is also estimated how many interaction partners are in the scenario and where they are.

1.1. Challenges of CAVSA

CAVSA is a challenging task. Let us look at a scenario to illustrate this. Assume a robot is working behind a bar and dealing with more than one guest. A typical bar scene is shown in

Figure 1.1.: A typical bar scene with more than one guest. Image originated from: http://pintspub.com/.

Fig. 1.1. These guests have never been seen and heard before. They dynamically come to the bar for the orders and leave after getting their drinks.

In order to fulfill the tasks as a bartender, the robot should at least understand the scene and be able to answer the following questions:

- How many people are in the surroundings? Sometimes it is not easy to estimate the number. First, guests dynamically come to the bar and leave, so that the number of guests is changing over time. Second, the robot can not see all the people at the same time due to limited camera field of view (FOV) and head movements.

- What is the location of the guests? Localization of people is important in order to put the drinks in front of a guest. The guests may move all the time and even change their locations outside the camera FOV, the robot should be able to track the guests.

- Who is speaking? The robot should assign words to the corresponding speakers. Otherwise, it could give a drink to the wrong guest. Additionally, when a guest has ordered a drink, leaves the bar for a while and returns to pick his drink, the robot should recognize him again. Moreover, a sound could be related to a person that stands outside the camera FOV and has not been seen before.

In the following, the contributions of this thesis and strategies to overcome the above challenges are outlined.

1.2. Contribution of this Thesis

The research goal of this thesis is to develop a system for CAVSA with a focus on human-robot dialogues in multi-person environments. The main contribution lies in the **system architecture** as illustrated in Fig. 2.1. Three important concepts are applied together in the system: proto-object, Short Term Memory (STM) and audiovisual association.

- **Proto-Objects:** Most previous approaches in robotics such as [Okuno et al., 2004] rely on low-level position estimations, i.e. the position of a face in the current image or the computed sound position in a 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, the concept of proto-objects, a mid-level scene representation between pixel (or voxel) and semantic object level, has been proposed [Bolder et al., 2007, Rodemann et al., 2009a]. 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). Thus, proto-objects can be tracked, pointed or referred to without identification and enable a flexible interface to behavior-control in robotics [Bolder et al., 2007]. In this thesis, proto-objects are used as a compressed collection of a variety of features. They can basically combine an arbitrary number of features and can be easily extended with new features to improve performance without changing the architecture of the system. In the past, visual [Bolder et al., 2007] and audio [Rodemann et al., 2009a] proto-objects were used separately. The CAVSA system combines the two approaches into one consistent framework.

- **Short Term Memory (STM):** As the robot usually can not see all the interaction partners at the same time due to limited camera FOV and head movements, STM can be used for the representation of the whole scene (see e.g. [Bolder et al., 2008, Schmüdderich, 2010]). Furthermore, STM is necessary for cognitive tasks. For example, when a person leaves the scene for a while and returns, the robot should recognize him again. Additionally, STM can be used to keep track of environmental dynamics such as changes in the number and position of speakers, and acquire knowledge of new speakers continuously in an unsupervised manner.

- **Audiovisual association:** This work combines visual and auditory modalities to achieve better performance. The integration of auditory and visual information derived from the same event can enhance the accuracy of the resulting perception [Bulkin and Groh,

2006]. Moreover, auditory and visual modality are somehow complementary for scene representation [Bulkin and Groh, 2006]. On the one hand, the vision channel transmits more accurate position information than the auditory channel in typical situations. On the other hand, while visual cameras usually have a limited FOV, auditory perception is possible in 360°.

The CAVSA system links the current auditory signal to the matched visual signal to answer the question "Who is speaking?". The audiovisual association is based on either temporal coincidence (synchrony) or temporal and spatial coincidence depending on the number of visual sources in the scene. The uncertainty of the audiovisual association is also measured. Additionally, the situation is considered where a sound is related to a person who has not yet been seen.

Moreover, since the speakers have not been seen or heard before, it is impossible to train high-level auditory and visual features beforehand to recognize voice and face. Hence the strategy is to **combine many simple auditory and visual features to represent speakers**. Experiments from neuroscience could show that the recognition of sensory inputs involves the interaction of different brain circuits, each of which is activated by a certain type of feature [Wang and Brown, 2006]. Humans normally recognize an object with more than one visual feature such as color, shape and size. The brain binds different features together to represent the input signal. From a computational perspective, it is also advantageous to combine multiple features of the same source, because multiple features may improve the accuracy using redundant information. A system using only a single feature lacks robustness, it easily fails when the feature is missing or observations are ambiguous. The CAVSA system combines many simple auditory and visual features to differentiate between speakers, although each individual feature is normally not good enough to fulfill this task.

1.3. Applications

CAVSA has many potential applications. This thesis assumes that all meaningful sounds originate from human speakers and applies CAVSA in two applications.

The first application is online adaptation of audio-motor maps. Audio-motor maps describe the relationship between binaural audio cues and sound position in motor coordinates (azimuth and elevation). These audio cues such as interaural time difference (ITD) and in-

teraural intensity difference (IID), result from the interaction of the head and ears with the incoming auditory stimulus [Wang and Brown, 2006]. Using audio-motor maps, a binaural auditory system can obtain sound source positions from measured audio cues. More details on audio-motor maps can be found in Appendix B. Since vision plays an important role in calibration of audio-motor maps in humans and animals [Zwiers et al., 2001, Knudsen, 2002; 1998], it is used as the feedback signal for precise position information. It is then necessary to match a visual signal to the current sound, which is challenging when more than one visual source exists. If an unrelated visual signal is selected for the adaptation, the quality of audio-motor maps can deteriorate. Therefore, a fundamental problem in online adaptation of audio-motor maps is audiovisual association in scenarios where the current acoustic signal is to be related to one out of many visual signals (e.g. hearing an utterance and seeing multiple faces at the same time). State of the art approaches [Nakashima and Mukai, 2005, Hörnstein et al., 2006] employ heuristics for linking visual and auditory information and can only work in constrained environments where only one face or one colored marker appears. It will be shown that CAVSA is able to find the correct visual correspondence of the current sound source and enable online adaptation to run in more complex environments. Furthermore, the quality of audio-motor maps depends on the performance of CAVSA, given precise measurements of visual position and audio cues. This is the reason why CAVSA is tested in online adaptation of audio-motor maps. In this application, a small set of simple features is sufficient, since it is not necessary to estimate person IDs.

CAVSA is also tested in human-robot dialogue scenarios. In a dialogue scenario, it may not be so important to figure out exactly who a person is, but to notice when the camera sees the same person again after the robot head has moved. The robot should also be able to recognize people even when they changed position while out of the camera FOV. This task requires a bigger set of visual features compared to the online adaptation of audio-motor maps. Assuming that people have not been seen and heard before, a database for training high-level features beforehand to recognize faces is not available. Hence the system combines a set of simple visual features such as height, color and texture of different upper body parts. It will be shown that a system integrating these simple visual features, is able to deal with complicated situations. For example, speakers can be recognized again when they leave and enter the scene.

Furthermore, a set of simple auditory features is tested to recognize voices and environmental sounds. It will be shown that auditory features have good performance to recognize environmental sounds and thus are used for filtering these sounds.

1.4. Structure

The remainder of this work is structured as follows: In the following chapter the system architecture of CAVSA, especially the three important concepts: proto-object, STM and audiovisual association, are introduced. Chapter 3 details the application of CAVSA in online adaptation of audio-motor maps and presents the results. In chapter 4, more visual and auditory features are introduced and tested for the application in human-robot dialogue scenarios. Chapter 5 finally summarizes the work and provides suggestions for future work.

2. System Architecture and Components

As mentioned in the previous chapter the main contribution of this work lies in the system architecture. This chapter first introduces the system architecture of CAVSA in section 2.1 and then three main components of the system in section 2.2-2.4. Section 2.2 introduces visual and audio proto-objects. Section 2.3 explains the procedure of inserting an incoming proto-object into the Short Term Memory (STM). Section 2.4 reviews the literature in audiovisual association, describes how to associate audio and visual proto-objects in case of multiple visual proto-objects, and how to calculate the uncertainty of the audiovisual association. Finally, two applications of CAVSA are shown in section 2.5. Important aspects of the system architecture proposed in this chapter were already published in [Yan et al., 2011a; 2013a].

2.1. Architecture

The CAVSA approach proposed in this work focuses on the analysis of a dialogue scenario with an unknown number of speakers who have not been seen or heard before. Its main goals are the estimation of number and position of speakers as well as the identification of the current speaker. This work intentionally ignores the issue of sound source separation, because there exists already a large body of research on the topic [Brown and Wang, 2005, Pedersen et al., 2008]. In the following it is assumed that all speakers talk in sequence or sound source separation has already been performed and that background noise is stationary. The stationary background noise can be largely removed using standard methods such as spectral subtraction [Rodemann et al., 2006a].

The system first combines several auditory and visual features into proto-objects. Proto-

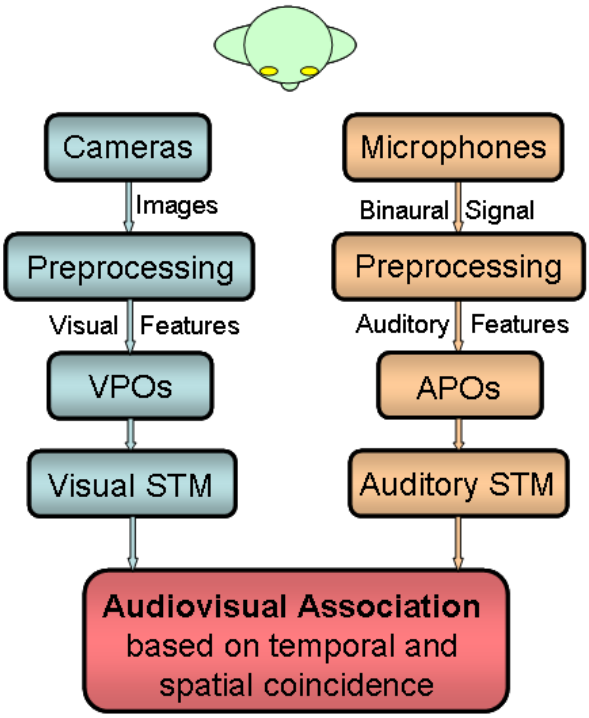


Figure 2.1.: System architecture of Computational AudioVisual Scene Analysis (CAVSA). APO: audio proto-object, VPO: visual proto-object. In CAVSA the scene is represented with audio and visual proto-objects. Audio and visual Proto-objects for the same speaker are then grouped together in their STMs respectively. Finally, audio and visual proto-objects are matched based mainly on temporal and spatial coincidence.

objects are designed for operation in realistic environments, since they can be tracked, pointed or referred to without identification and offer a flexible interface for behavior-control in robotics. In the past, both visual and audio proto-objects were used separately [Bolder et al., 2007, Rodemann et al., 2009a]. In this work the two approaches are combined into one consistent framework. In this manner, dense visual images and sparse sound sources are transformed into a common representation, which eases audiovisual integration. Proto-objects for the same speaker are then grouped together in auditory STM and visual STM, respectively. One advantage of using STM is that it allows the system to memorize persons who disappear for a while. Finally, visual and audio proto-objects of the same speaker are associated. The current sound is linked to the corresponding speaker in multi-person environments, and the uncertainty of the audiovisual association is measured. Fig. 2.1 schematically illustrates the system architecture.

The three main components of the CAVSA system: proto-objects, STM and audiovisual association, are described in the following sections.

2.2. Proto-Objects

Low level sensory representations are often not well suited to guide behavior in real-world environments, due to rapid changes of the signals, measurement noise and a high complexity of scenes. Therefore, this thesis uses proto-objects as an intermediate level feature representation to build a more stable and robust representation of the environment.

The term "proto-object" is inspired by earlier work on the representation of sensory information in the brain [Burr and Alais, 2006, Hershey and Movellan, 1999, Shinn-Cunningham, 2008]. 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 the pixel (or voxel) level and a semantic object level [Schmüdderich, 2010, Bolder et al., 2007, Rodemann et al., 2009a].

Visual proto-objects have been widely used in cognitive robotics, for example to represent behaviorally relevant objects, object parts or groups of objects in the environment [Pylyshyn, 2001, Schmüdderich, 2010, Bolder et al., 2008]. In [Rodemann et al., 2009a], proto-objects are transferred to the auditory domain as a mid-level representation of sounds. While the concept of visual proto-objects as it is used in most of the current work refers to little more than a localized feature and includes some but not all of the properties of objects [Pylyshyn, 2001], Rodemann et.al [Rodemann et al., 2009a] use audio proto-objects as a compressed representation container which can combine an arbitrary number of features. In this manner, new features can be added easily without changing the system architecture. In this thesis, both audio and visual proto-objects in the same structure as in [Rodemann et al., 2009a] are used to technically ease audiovisual integration.

In the visual domain, regions (x/y in the camera image) are segmented based on homogeneity, edges or object detection. A typical visual proto-object would contain a position (for example the 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 [Bolder et al., 2007]. In the audio domain, a simple threshold-based segmentation process similar to Voice Activity Detection (VAD) [Ramírez et al., 2007] is used to separate sound pieces from the background. A segment is a perception unit, which can be one or more words, or a whole sentence. 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 segment (such as the

start time), the segment length and averaged audio features (such as the mean estimated position over the complete length of the sound, histograms of localization features like IID and ITD, and the mean energy of the sound). Which features are collected in proto-objects is application dependent.

As can be seen from the above, there are many benefits of using 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 [Bolder et al., 2007].

2.3. Short Term Memory

In biology, a short term memory contains the most recently perceived items which can be lost with passage of time [Miller, 1956, Atkinson and Shiffrin, 1971]. Auditory and visual STM work separately and similarly [Visscher et al., 2007]. Visual STM can maintain continuity across visual interruptions caused by eye movements, eye blinks, and other visual interruptions. Especially, saccadic eye movements occur 3 times per second and interrupt the visual processing briefly [Hollingworth and Henderson, 2002]. Visual STM retains and combines visual representations over one or more saccadic eye movements. Furthermore, since seen regions may be different over time due to eye's or head's movements, visual STM can construct a larger-scale representation of the whole scene by combining previous representations [Hollingworth and Henderson, 2002]. In comparison, auditory STM maintains the representations of recently heard sounds for further cognitive tasks such as speech recognition and voice recognition [Yost et al., 2008]. To summarize, auditory and visual STM are essential for scene representation in biology. Inspired by this, they have been built in robotics [Bolder et al., 2008, Schmüdderich, 2010, Markov, 2009, Bennewitz et al., 2005].

Within the CAVSA system, STM is used for the scene representation similar to [Schmüdderich, 2010]. As shown in Fig. 2.2, the procedure of inserting an incoming proto-object into STM can be described as follows:

- "*compare*" step: It is decided whether the incoming proto-object belongs to one of

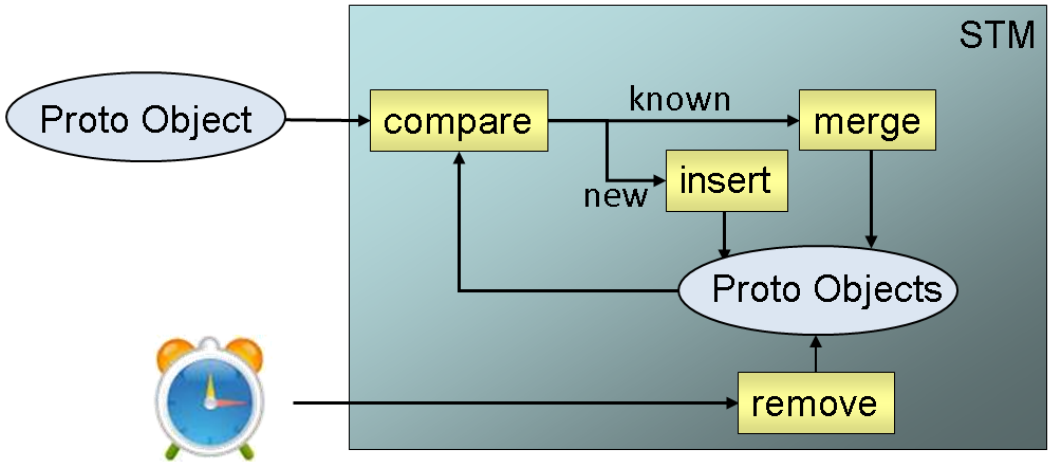


Figure 2.2.: The procedure of inserting an incoming proto-object into STM.

the known speakers or a new speaker. The distance or similarity of selected grouping features is computed between the incoming proto-object and all proto-objects in the STM. The calculation of distance and similarity is explained in the following chapters. After distance calculation, the distance between the incoming proto-object and the closest proto-object in the STM is compared with a threshold Θ_S. If the distance is smaller than Θ_S, they are assumed to belong to the same known speaker. Otherwise, the incoming proto-object is assumed to belong to a new person.

- "*insert*" step: If the incoming proto-object is assumed to belong to a new person, the proto-object is inserted into the STM. In order to cope with the spurious detection of proto-objects, status flags are used [Schmüdderich, 2010]. The flag of a new inserted proto-object is set to "volatile".

- "*merge*" step: If the incoming proto-object is assumed to belong to a known speaker, it will be merged with the closest proto-object in STM. Furthermore, if a "volatile" proto-object is re-detected in a given period T_{in}, the flag is changed to "found" and the proto-object becomes valid. Since a spurious detection due to noisy input occurs only for a short time, T_{in} is set to two update intervals, i.e. at least two immediate detections are necessary for a valid proto-object [Schmüdderich, 2010].

  The merge operation is done by a low pass filter:

$$y(t) = y(t-1) \cdot w + x_{in}(t) \cdot (1-w), \tag{2.1}$$

  Here, y is one of the grouping features in the closest proto-object in the STM, x_{in} is the

corresponding feature in the incoming proto-object and t is the current time step. In this manner, all features in these two proto-objects are correspondingly merged. The weight w is defined as an activity function:

$$w = e^{-\frac{t_k}{\tau}}, \tag{2.2}$$

where τ is a decay factor and t_k stands for the time in seconds since the closest proto-object has been updated the last time. If the proto-object is currently updated, t_k is set to 0. w indicates the reliability of features in the closest proto-object that was updated in the recent time step. The longer a proto-object has not been seen, the less reliable its features become. The value of τ is set in accordance with the applications.

- "*remove*" step: "Volatile" proto-objects, which are not updated for more than T_{in} are removed from the STM. In this manner, falsely detected proto-objects are quickly deleted. "Found" proto-objects, which are not updated for more than a certain period T are also removed. With the memorization mechanism, missing detections of proto-objects due to noisy input or occlusion are tolerated. If T is too small, the STM can not remember a person who leaves or is occluded for a while. Conversely, if T is too large, STM may contain too many proto-objects and is not able to dynamically represent the current scene, e.g. after a person has already left the scene, his corresponding proto-object is stored in the STM for a long time. T is set to 100 s in the experiments.

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 case of visual STM, it is noticed that when more than one person appears in the camera image and more than one proto-object is stored in visual STM, two or more incoming proto-objects may select the same "most similar" proto-object in STM. To avoid the selection confusion, a simple greedy mapping method is used. Let X_m $(m \in [1,M])$ denote the m-th current incoming proto-object, Y_n $(n \in [1,N_V])$ denote the n-th proto-object in STM, where M represents the number of current incoming proto-objects and N_V the number of proto-objects in STM. If d_{mn} stands for the distance between X_m and Y_n, all distances d_{mn} build a matrix D' with M rows and N_V columns. For the greedy algorithm as described in Algorithm 1, 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. This algorithm is similar to the one used in [Hung and Friedl, 2008].

Algorithm 1 Greedy algorithm for the visual STM

1: build a $M \times N_V$- matrix D' with all d_{mn} $(m \in [1,M], n \in [1,N_V])$
2: search for $d_{m'n'} = \min d_{mn}$
3: initialize the number of eliminated rows: $M_E = 0$
4: **while** $d_{m'n'} < \Theta_S \wedge M_E < M$ **do**
5: merge corresponding features in proto-object $X_{m'}$ and $Y_{n'}$ as in Eq. 2.1
6: eliminate the m'-th row and the n'-th column in matrix D'
7: $M_E \leftarrow M_E + 1$
8: search for $d_{m'n'} = \min d_{mn}$
9: **end while**
10: **if** $M_E < M$ **then**
11: insert the unassigned incoming proto-objects into STM
12: **end if**

2.4. Audiovisual Association

The research field of multimodal HRI, especially based on auditory and visual modalities, is quickly developing. Multimodal HRI makes human-robot interfaces appear more natural compared to those using only a single modality. For instance, lip movements extracted from camera images are used to improve speech recognition performance of the robot [Iwano et al., 2007]; pointing gesture sensed through cameras helps the robot better understand spoken languages [Bui, 2006]. In this work, visual and auditory modalities are integrated to enhance the scene representation.

As shown in Fig. 2.3, the first step of audiovisual integration is to associate an auditory and a visual signal that are caused by the same event. The current audio proto-object is linked to a visual proto-object based on either temporal or temporal and spatial coincidence, depending on the number of visual proto-objects. Temporal coincidence (synchrony) has been identified as one of the most important factors determining whether or not multisensory integration takes place [Noesselt et al., 2007]. 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 [Alais and Burr, 2004]. If more than one source exists, other information such as position information is needed to avoid ambiguity. The closer a visual

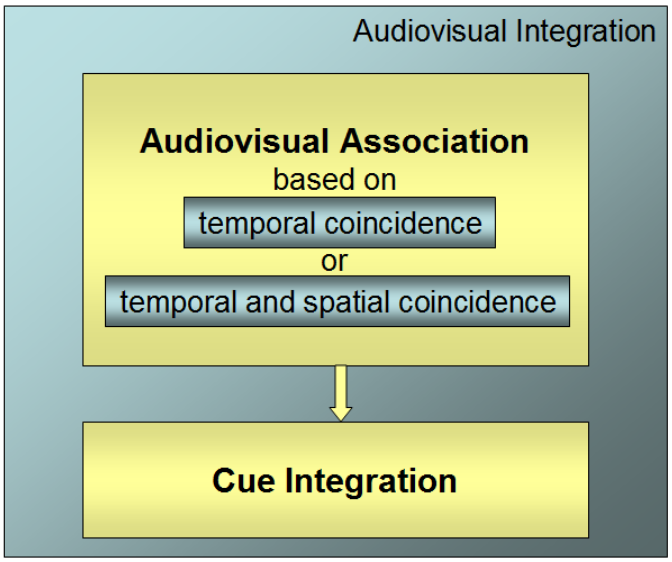


Figure 2.3.: The procedure of audiovisual integration. The first step is to associate an auditory and a visual signal that are caused by the same event. The audiovisual association is based on either temporal or temporal and spatial coincidence. The second step is to integrate auditory and visual cues, that are auditory and visual position in this context, caused by the same source.

stimulus is to an auditory stimulus, the more probable they are perceived as having a common cause [Wallace et al., 2004]. 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 [Knudsen and Knudsen, 1985, Wallace and Stein, 2006]. When the animals are raised in artificial environments where auditory and visual stimuli are temporally coupled but spatially incoherent, multisensory neurons in the superior colliculus (SC) will still learn 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 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 person in the scene). Otherwise, when more than one face appears in the scene, audiovisual integration is based on both temporal and spatial coincidence. The second step of audiovisual integration is to integrate multisensory cues, that are auditory and visual position in this context, caused by the same source. This work focuses on the first step and simplifies the second step by considering the visual position as the position of the speaker, since visual position is normally more accurate than auditory. Readers interested in cue integration find e.g. integration of auditory and visual position information in [Weisswange et al., 2011].

In the following, an overview of related work in audiovisual association is first presented. Then it is explained how to associate audio and visual proto-objects based on probabilities in case of multiple visual proto-objects, as well as how to calculate the uncertainty of the

audiovisual association.

2.4.1. Related Work in Audiovisual Association

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) [Hershey and Movellan, 1999, Hung and Friedl, 2008, Noulas and Krose, 2007, Casanovas et al., 2007, Friedland et al., 2009, Fisher and Darrell, 2004]. 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 [Kim et al., 2007, Checka et al., 2003, Khalidov et al., 2008, Nakadai et al., 2001]. 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. While the robot head is moving, a speaker may leave the field of view (FOV) for a while and reappear again, as most robots have a limited FOV. In this case, most of these methods such as [Kim et al., 2007, Checka et al., 2003, Khalidov et al., 2008] do not memorize the speaker and would consider him as a new person. In contrast, [Nakadai et al., 2001] can recognize persons using face recognition, but the training of faces is required beforehand. In contrast to these methods, the audiovisual association in this work is based on either temporal or temporal and spatial coincidence, depending on the number of visual sources as mentioned. The uncertainty of audiovisual association is also computed. Compared to [Nakadai et al., 2001], the database for training high level features such as features for face recognition is not available, as the CAVSA system deals with scenarios with an unknown number of speakers who have not been seen or heard before. Instead, STM is used to memorize a person who disappears briefly.

2.4.2. Probability Based Audiovisual Association

In the following it is explained how audio and visual proto-objects are associated in case of multiple visual proto-objects. First, the probability that an audio proto-object and a visual proto-object have a common cause is calculated based on their position information. The closer a visual proto-object is to an audio proto-object, the more probable they are caused by the same person.

Let A_i ($i \in [1, N_A]$) and V_j ($j \in [1, N_V]$) denote an audio proto-object in auditory STM and a visual proto-object in visual STM, respectively. N_A stands for the number of proto-objects in auditory STM, and N_V the number of proto-objects in visual STM. The relative probability that audio proto-object A_i and visual proto-object V_j have a common cause is denoted as $P_{common}(A_i, V_j)$. The calculation of $P_{common}(A_i, V_j)$ will be described later in Eq. 3.1 and Eq. 4.11 for concrete applications.

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

$$l = \underset{j \in [1, N_V]}{\arg\max}\, P_{common}(A_i, V_j) \tag{2.3}$$

If there is only one visual proto-object ($N_V = 1$), it is treated as the current sound source. That is, only **temporal coincidence** of proto-objects is employed for audiovisual association.

2.4.3. Uncertainty of Audiovisual Association

Dependent on applications, it may be necessary to check how reliable the association of the current audio proto-object and the visual proto-object with the maximum probability is. This work calculates the uncertainty of the audiovisual association using the Shannon informational entropy [Shannon, 1948] of the set of normalized probabilities. Shannon informational entropy is well-known and has been widely used to measure the uncertainty of probability distributions, for example in speech recognition [Heckmann et al., 2002].

If one visual proto-object shows a very high probability and all other visual proto-objects have low probabilities, this expresses a low entropy indicating a reliable association. Conversely, when all visual proto-objects have quasi equal probability, the entropy is high and

the association is unreliable. The normalized probability is computed as

$$\widehat{P}_{common}(A_i,V_j) = \frac{P_{common}(A_i,V_j)}{\sum_{i=1}^{N_V} P_{common}(A_i,V_j)}, \tag{2.4}$$

and the entropy for A_i is given as

$$H_i = \begin{cases} 0 & \text{if } N_V = 1, \\ \frac{-\sum_{j=1}^{N_V} \widehat{P}_{common}(A_i,V_j)\cdot\log_2 \widehat{P}_{common}(A_i,V_j)}{\log_2 N_V} & \text{if } N_V > 1. \end{cases} \tag{2.5}$$

Here, the division by $\log_2 N_V$ ensures that the maximal H_i is 1 to easily set a threshold Θ_H which is dependent on the applications. The smaller H_i, the more reliable is the association of the audio proto-object A_i and the visual proto-object with the maximum probability. Note that, in case of only one visual proto-object ($N_V = 1$), which is treated as the current sound source due to temporal coincidence, the audiovisual association is reliable ($H_i = 0$). The uncertainty of the whole audiovisual association can be captured by averaging over all H_i:

$$H = \frac{\sum_{i=1}^{N_A} H_i}{N_A}. \tag{2.6}$$

2.5. Overview of Applications

In the following chapters, CAVSA is first applied in online adaptation of audio-motor maps (chapter 3) to overcome the issue of audiovisual association in scenarios where the current acoustic signal is to be related to one out of many visual signals, for instance we hear an utterance and see many faces. Only a small set of simple features is used in the original version of CAVSA, since it is not necessary to identify persons in this application. Then CAVSA involving more complicated functions is used in human-robot dialogue scenarios (chapter 4). It aims to learn the number and position of speakers, as well as who is currently speaking. In this application, speakers should be recognized again e.g. when they leave and enter the scene. Thus the system needs to integrate more visual and auditory features to deal with such complicated situations. All features are stored in the same format, a biologically-inspired population coding [Pouget et al., 2000] that combines a flexible format with dense information related to probabilities. Thanks to the common structure of different features, proto-objects can have a variable number of these features, so that new features can be added

easily. Moreover, since audio and visual cues have the same structure, an audiovisual integration of arbitrary cues is technically feasible.

3. CAVSA in Online Adaptation of Audio-Motor Maps

In the previous chapter, the system architecture of CAVSA was proposed. This chapter deals with the first application: online adaptation of audio-motor maps.

While the visual system projects directly a point in the outer world onto the camera image, the auditory system relies on the processing of implicit cues to derive the sound source position. Given audio cues, a binaural auditory system estimates sound source positions using audio-motor maps, which represent the relationship between audio cues and the position of the sound source. More information on audio-motor maps is explained in Appendix B. Audio-motor maps can be calibrated offline by measuring audio cues for several known positions [Finger et al., 2010]. However, audio-motor maps can change and need to be relearned whenever any relevant part of the robot or the environment is modified, for example, microphone type, microphone position, robot head and room. Additionally, since it is difficult to estimate the quality of the current maps, making the decision whether to relearn the current maps is hard. Hence a continuous online adaptation is essential for a mobile robot.

Since vision plays an important role in calibration of audio-motor maps in humans and animals [Zwiers et al., 2001, Knudsen, 2002; 1998], it is used as the feedback signal for precise position information in online adaptation. It is then necessary to match a visual signal to the current sound, which is challenging when more than one visual source exists. If an unrelated visual signal is selected for the adaptation, the quality of audio-motor maps can deteriorate. Therefore, a fundamental problem in online adaptation of audio-motor maps is audiovisual association in scenarios where the current acoustic signal is to be related to one out of many visual signals (e.g. hearing an utterance and seeing multiple faces at the same time). CAVSA can be applied to solve this problem. It searches for the correct visual correspondence of the current sound source and enables online adaptation to run in free interaction with a number of a-priori unknown speakers.

It is interesting to see if the system is able to learn the audio-motor maps from a random initialization in different scenarios, if it is robust in dynamically changing scenarios with multiple people, 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, the system computes the difference between online-adapted and offline-calibrated audio-motor maps and also compares sound localization results with ground truth data (if available).

The outline of this chapter is as follows: Section 3.1 shows drawbacks of the related work in learning of audio-motor maps, that is the motivation of using CAVSA. Then the setting of CAVSA for online adaptation is introduced in section 3.2, and section 3.3 explains how to adapt audio-motor maps. The results of testing the bootstrapping ability, robustness and adaptation ability are shown in section 3.4. Finally, a discussion is given in section 3.5.

Important aspects of this application and the corresponding results were already published in [Yan et al., 2011a;b; 2013a].

3.1. 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. Willert et. al. [Willert et al., 2006] 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 [Knudsen and Knudsen, 1989, King et al., 1988, Hyde and Knudsen, 2002, Paul M. Hofman, 1998, Gold and Knudsen, 2000, Kacelnik et al., 2006, DeBello and Knudsen, 2004]. 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.

[Rodemann et al., 2007] 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 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 [Zwiers et al., 2001, Knudsen, 2002; 1998], several recent papers [Hörnstein et al., 2006, Nakashima and Mukai, 2005] suggest to use visual feedback to provide the sound source position. These approaches are very appealing but are demonstrated only in simple scenarios. [Nakashima and Mukai, 2005] attached a red color marker on the loudspeaker to indicate the source, while [Hörnstein et al., 2006] 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.

These approaches [Rodemann et al., 2007, Nakashima and Mukai, 2005, Hörnstein et al., 2006] 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 in interaction with the real-world environment. Animals and humans do not have a reliable calibration phase and their head is growing continuously during development. How can an autonomous agent acting in a real-world environment adapt its audio-motor map like animals and humans? 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. This system also uses vision as the supporting sensory modality for the online adaptation of audio-motor maps.

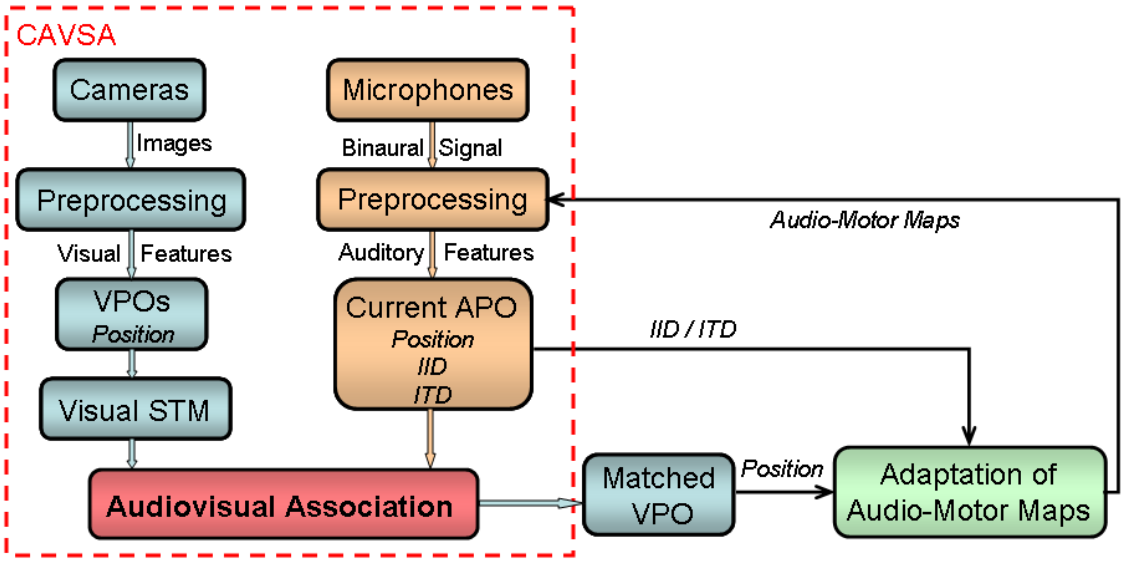


Figure 3.1.: System architecture of CAVSA in online adaptation of audio-motor maps. CAVSA applies the current audio-motor maps for sound localization and in turn supports online adaptation of audio-motor maps by searching for the matched visual proto-object of the current sound. APO: audio proto-object, VPO: visual proto-object.

In contrast, the system explicitly tackles the issue of the proper alignment of audio and visual sources through CAVSA. Additionally, the 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.

3.2. Setting of CAVSA for Online Adaptation

As shown in Fig. 3.1, 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.

This work concentrates on audio-motor maps for azimuth to ease the description of the algorithm, but the approach can be expanded to elevation as well. Azimuth angles between -90° and 90° are taken into account because of the mechanical constraint of the robot head. Additionally, since the mapping between audio cues and sound source location is smooth for a typical robot design, it suffices to know the mapping for a couple of positions and interpolate intermediate locations. It is found that a grid with a spacing of 10° in azimuth angle is well suited for practical applications.

In this work, the audio-motor map is represented as a simple look-up table $M^i_\theta(f,n)$ ($i = 1$ 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 θ. An example of an IID map $M^1_{-90}(f,n)$ at -90° is shown in Fig. C.2 (in Appendix C). Given audio-motor map $M^i_\theta(f,n)$ and measured cues $C'^i(f,n)$, localization amounts to finding the position θ with the highest overlap with measured cues. For more information about $M^i_\theta(f,n)$ and $C'^i(f,n)$ see Appendix C. 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.

CAVSA is used to search for the corresponding visual part of the current sound in online adaptation of audio-motor maps. A small set of simple features are sufficient for this application. 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). Moreover, auditory STM is not used, since only information about the current sound is required in online adaptation of audio-motor maps. Additionally, the current sound is sometimes related to a person who has not yet been seen. This situation is considered as well. As shown in Fig. 3.1, the current audio proto-object is used together with the visual STM in the CAVSA to search for the matched visual proto-object. The extraction of auditory and visual features, as well as the configuration of visual STM and audiovisual association are introduced in the following.

3.2.1. Audio Processing

The system employs a biologically-inspired audio processing system using two microphones on a humanoid robot head. As shown in Fig. 3.2, the first stage in audio-processing is a Gammatone Filterbank (GFB), which is considered to be a good model of the human cochlea [Slaney, 1993]. 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 the signal envelope is extracted for every frequency band. In order to deal with stationary background noise, the system employs a spectral-subtraction technique that estimates the level of noise in each frequency channel and then reduces this noise [Rodemann et al., 2006a]. This method is very effective against noise sources like air-conditioning or computer fan noise. Thereafter, signal energy is calculated as

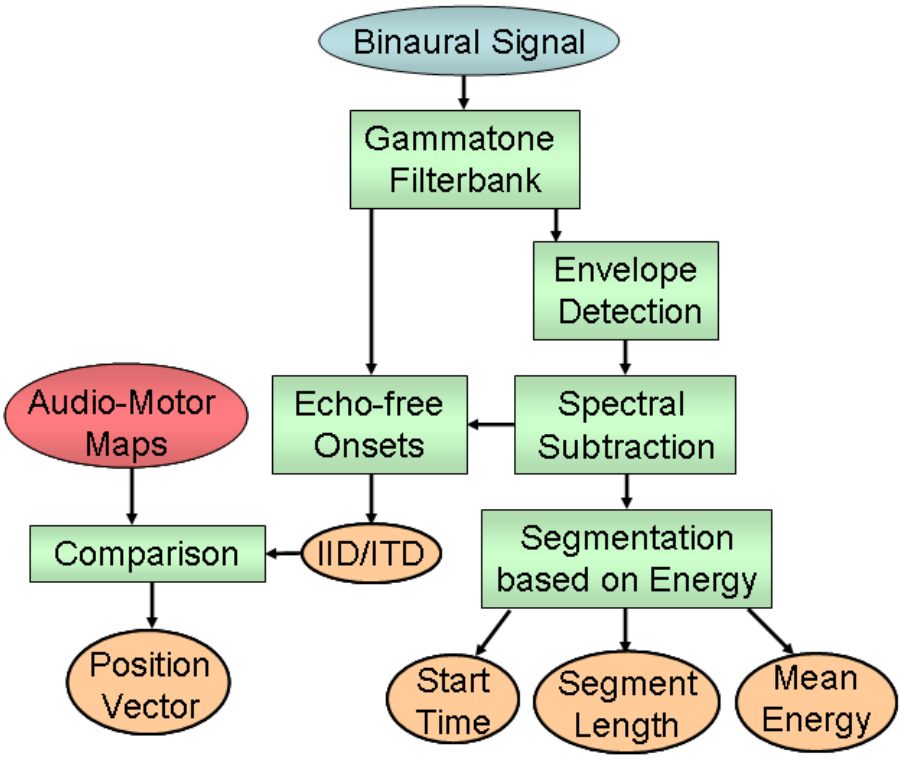


Figure 3.2.: Audio processing. For online adaptation of audio-motor maps, an audio proto-object contains start time, segment length, the mean energy over all samples of the segment, position evidence vector and audio cues (IID and ITD).

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, start time, length and the mean energy over all samples of the segment are derived and saved 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. The system uses a biologically inspired strategy, the precedence effect, to reduce the impact of echoes [Heckmann et al., 2006]. The basic principle is the detection of signal onsets, which are quick and strong changes in the frequency channel's envelope. The system only measures audio cues IID and ITD for sound localization during the signal onsets (before echoes arrive). [Rodemann et al., 2006a, Heckmann et al., 2006, Rodemann et al., 2008] have shown that such a sound localization system is able to efficiently operate even in noisy and echoic environments. In this thesis, population-coded audio cues are used for sound localization, for more information see Appendix C.

Given audio-motor map $M^i_\theta(f,n)$ and population-coded cues $C'^i(f,n)$, $C'^i(f,n)$ is compared

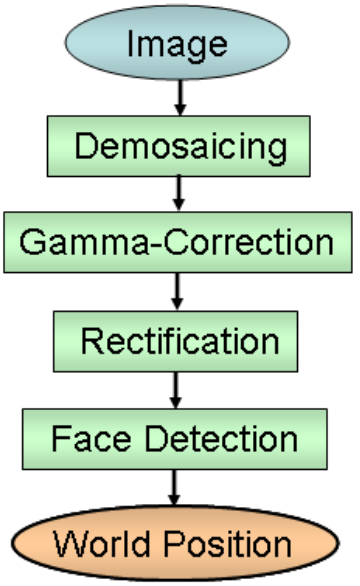


Figure 3.3.: Visual processing

with $M_\theta^i(f,n)$ for all positions θ by computing scalar products to acquire the position evidence vector W_θ^i. The peak in the position vector W_θ ($W_\theta = \sum_{i=1}^{2} W_\theta^i$) is taken as the estimated sound source position. Population-coded cues $C'^i(f,n)$ and the position evidence vector W_θ are saved 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 typical background sounds are rather short and have less energy than robot directed calls. Filtering of background sounds is explained in [Rodemann et al., 2009b].

3.2.2. Visual Processing

The robot head is equipped with a set of stereo color cameras. As shown in Fig 3.3, 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, visual features of a person are extracted to form a visual proto-object. In the camera field of view (FOV), the presence of persons is detected by searching for their faces. It is assumed that persons talking to the robot look at its face most of the time. Face detection is a low-level mechanism. It was found that human newborns have already inborn face detection mechanisms [Stein et al., 2011, Farroni et al., 2005]. A frontal face detection algorithm developed by Viola and Jones [Viola and Jones, 2001] 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. Since the system

does not rely on visual distance estimation, only a single camera is used. The distance between the robot head and the speakers is about 1 m. Using the distance information and saccade maps (see [Rodemann et al., 2006b]), the system converts camera positions into 3D world coordinates, and stores them as well as the time of the current image in the visual proto-object. Instead of this fixed setting, a lot of algorithms for depth estimation using stereo cameras could be applied to obtain 3D positions.

3.2.3. Visual STM

Since the robot has a limited FOV, a speaker may leave the FOV e.g. due to movements of the robot head. In this case, visual STM can still memorize his related proto-object. The procedure of entering an incoming proto-object into the STM is introduced in section 2.3. Initially the STM is empty. When a new visual proto-object appears, it is copied into the STM. Unless it is updated, it will disappear after a certain time-out T, which is set to 100 s in the experiments. The system determines if an incoming proto-object is new or corresponds to a proto-object in the STM, based on the Euclidean distance of their positions in 3D world coordinates. The assumption is that people do not change their position much in a dialogue scenario, so that the position difference of the same person is small between two consecutive views. When an incoming proto-object X_m $(m \in [1,M])$ appears, it will be compared with each proto-object Y_n $(n \in [1,N_V])$ already in the STM. M and N_V denote the number of incoming proto-objects and the number of proto-objects in the visual STM, respectively. The Euclidean distance of their 3D positions, represented as d_{mn}, is calculated. Then Algorithm 1 is used to merge each incoming proto-object with a stored proto-object or insert it into the STM. Threshold Θ_S of position difference is set to 0.3 m, so that slight movements of speakers such as head shaking are tolerated. Additionally, the parameter τ for the merge operation, in Eq. 2.2, is set to 5 s based on empirical tests.

3.2.4. Audiovisual Association

Audio and visual proto-objects are linked using their position information as mentioned in section 2.4. The closer a visual proto-object is to an audio proto-object, the more probable they are caused by the same person.

1) Basic approach

First, auditory position evidence vectors and visual positions in world coordinates have to be converted to the same metric, we choose motor coordinates (azimuth and elevation). This work concentrates 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 θ_a and θ_{v_j} denote the azimuth angle of the current audio proto-object A and visual proto-object V_j ($j \in [1, N_V]$), respectively. N_V stands for the number of visual proto-objects in visual STM. The relative probability that audio proto-object A and visual proto-object V_j have a common cause can then be approximated as:

$$P_{common}(A, V_j) = \exp\left(\frac{-|\theta_a - \theta_{v_j}|^2}{2 \cdot \delta^2}\right), \tag{3.1}$$

where the standard deviation δ is computed as the mean absolute difference in estimated azimuth between an audio and a visual proto-object which belong to the same speaker. δ 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:

$$\delta_s = \begin{cases} |\theta_a - \theta_v| \cdot w + \delta_{s-1} \cdot (1 - w) & \text{if } N_V = 1, \\ \delta_{s-1} & \text{otherwise.} \end{cases} \tag{3.2}$$

Here, s and w stand for update step and update factor respectively. We set $w = 0.1 \cdot \beta$ dependent on the fixed adaptation rate of audio-motor maps β, which controls the degree of adaptation for a single step (see also Eq. 3.11). If only one visual proto-object is in the STM, δ_s is updated. δ_0 is initialized as 40° to tolerate large position differences between audio and

Algorithm 2 Audiovisual association: Basic approach

1: **for** $j = 1$ to N_V **do**
2: Calculate $P_{common}(A, V_j)$ based on Eq. 3.1
3: **end for**
4: Search for V_l that has the maximum P_{common} as in Eq. 2.3
5: **for** $j = 1$ to N_V **do**
6: Calculate $\widehat{P}_{common}(A, V_j)$ based on Eq. 2.4
7: **end for**
8: Calculate entropy H as in Eq. 2.5
9: **if** $H \leqslant \Theta_H$ **then**
10: Associate A with V_l
11: **end if**
12: Update standard deviation δ as in Eq. 3.2

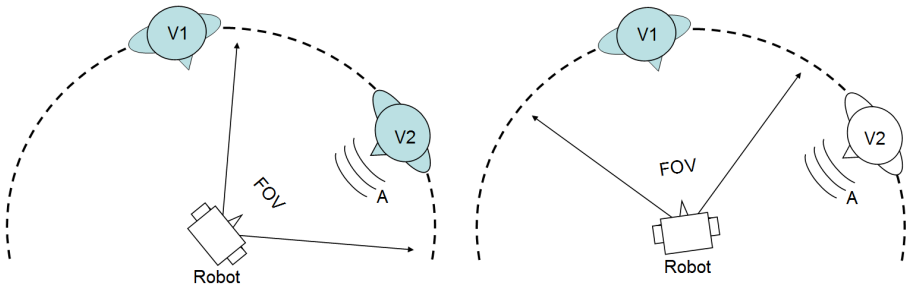


Figure 3.4.: Examples of two situations: *left*: all visual sources have been seen; *right*: the current sound source has not yet been seen. The head symbol with color means that the corresponding person has been seen. A is the current audio proto-object, V_1 and V_2 are visual proto-objects. The horizontal FOV is 90°.

visual proto-object when sound localization is very inaccurate at the beginning.

Then the visual proto-object V_l with the maximum probability P_{common} is searched for as in equation 2.3, and the uncertainty of the association of A and V_l is calculated as in equation 2.5. Θ_H is empirically set to 0.8. If entropy H is larger than Θ_H, A and V_l are not associated. The approach is described in Algorithm 2.

2) Consideration of unseen visual proto-objects

It is found that Algorithm 2 works well when all visual sources have been seen, but fails when the current sound source has never been seen. For ease of understanding, examples of the two situations are shown in Fig. 3.4. Therefore the system also considers the situation that the audio proto-object A is sometimes related to a visual proto-object that has not yet been seen [Yan et al., 2011b]. This is realized by using a view memory that memorizes how long a certain position has not been looked at. The view memory is denoted as $U(\theta)$, where θ is the azimuth angle in robot-centered coordinates. Each element in $U(\theta)$ is between 0 and 1. The larger the value $U(\theta)$ is, the longer the time since position θ has been last attended. $U(\theta)$ is described as

$$U(\theta) = 1 - e^{-\frac{t_\theta}{T_u}}, \tag{3.3}$$

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

For convenience, the unseen proto-object is denoted as V_{N+1}. Then the probability that audio

proto-object A is associated with V_{N+1} is described by:

$$P_{common}(A, V_{N+1}) = \frac{\sum_{\theta} U(\theta) \cdot P_{common}(A, V_{N+1}(\theta))}{\sum_{\theta} P_{common}(A, V_{N+1}(\theta))}, \tag{3.4}$$

where $V_{N+1}(\theta)$ denotes the unseen visual proto object that is assumed to be at position θ. If positions near the sound source have not been viewed or are not viewed for a long time, the related sound source has very probably not yet been seen. This leads to a large $P_{common}(A, V_{N+1})$. If $P_{common}(A, V_{N+1})$ is larger than the maximal $P_{common}(A, V_j)$ ($j \in [1, N_V]$), the matched visual signal is considered as not been seen so far. Otherwise, the association uncertainty is calculated as above using the entropy of normalized probabilities. The normalized probability and entropy become:

$$\widehat{P}_{common}(A, V_j) = \frac{P_{common}(A, V_j)}{\sum_{j=1}^{N_V+1} P_{common}(A, V_j)}, \tag{3.5}$$

and

$$H = \frac{-\sum_{j=1}^{N_V+1} \widehat{P}_{common}(A, V_j) \cdot \log_2 \widehat{P}_{common}(A, V_j)}{\log_2(N_V + 1)}. \tag{3.6}$$

We can see that if $N = 1$ and $P_{common}(A, V_{N+1}) \to 0$, then entropy $H \to 0$. This means that if only one visual signal appears, and positions near the current sound have been recently attended, the auditory and visual signals are assumed to have a common cause. Only in this case, standard deviation δ in Eq. 3.1 is updated. The update rule is described as below:

$$\delta_s = \begin{cases} |\theta_a - \theta_v| \cdot w + \delta_{s-1} \cdot (1 - w) & \text{if } N_V = 1 \wedge P_u < \Theta_{P_u}, \\ \delta_{s-1} & \text{otherwise.} \end{cases} \tag{3.7}$$

In comparison to Eq. 3.2, an additional condition $P_u < \Theta_{P_u}$ is added to $N_V = 1$. Here, $P_u = \widehat{P}_{common}(A, V_{N+1})$ is the normalized probability that A and V_{N+1} have a common cause, as described in Eq. 3.4. Θ_{P_u} is the threshold of P_u and is set to 0.1.

The method is summarized in Algorithm 3. The performance of Algorithm 2 and Algorithm 3 will be compared in section 3.4.3.

Algorithm 3 Audiovisual association: Consideration of unseen VPOs

for $j = 1$ to N_V **do**
    Calculate $P_{common}(A, V_j)$ based on Eq. 3.1
end for
Calculate $P_{common}(A, V_{N+1})$ as in Eq. 3.4
Search for V_l that satisfies
$$l = \underset{j \in [1, N_V+1]}{\arg\max}\ P_{common}(A, V_j) \wedge V_l \neq V_{N+1}$$
if $\exists\, V_l$ **then**
    for $j = 1$ to $N_V + 1$ **do**
        Calculate $\widehat{P}_{common}(A, V_j)$ based on Eq. 2.4
    end for
    Calculate entropy H as in Eq. 2.5
    if $H \leqslant \Theta_H$ **then**
        Associate A with V_l
    end if
    Update standard deviation δ as in Eq. 3.7
end if

3.3. Online Adaptation of Audio-Motor Maps

This section explains how to adapt audio-motor maps in consideration of the uncertainty of audiovisual association.

For online adaptation of audio-motor maps, the system needs to measure the IID/ITD of the current sound and estimate the corresponding visual position of the sound source. The matched visual proto-object is searched for by audiovisual association. The probability P_{common} (Eq. 3.1) for audiovisual association is based on the position difference. Hence, the system tends to choose a wrong visual proto-object when persons stand near the current speaker 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. 2.5) may refuse the audiovisual association. However, it was found in the experiments that wrong visual proto-objects in such cases can also enhance the quality of maps [Yan et al., 2011a], although they may have negative impacts in other applications. This is because of the smoothness of the mapping in audio-motor maps. Therefore, a new variable H' is defined to allow the use of these incorrect but useful visual proto-objects. The variable H' considers both entropy H and the position difference between the visual proto-objects with maximum and second maximum probability $\widehat{P}_{common}$:

$$H' = H \cdot |\theta_{v_1} - \theta_{v_2}|, \tag{3.8}$$

where θ_{v_1} and θ_{v_2} 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 difference 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, $\theta_{v_1} = \theta_{v_2}$ and $H = 0$.

The confidence c of an adaptation step is given by:

$$c = \begin{cases} 1 & \text{if } H \leqslant \Theta_H \vee H' \leqslant \Theta_{H'}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

In multi-person scenarios, the larger the threshold Θ_H ($\Theta_H \in [0,1]$) and $\Theta_{H'}$ are, the more matched visual proto-objects 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. Thresholds are chosen 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, audio-motor maps could be adapted when only one person is in the scenario. That is, both Θ_H and $\Theta_{H'}$ could be set to 0, so that online adaptation is stopped in multi-person scenarios. In contrast, in experiments with only several hours as in this work, 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 empirically set to 0.8, and $\Theta_{H'}$ adapts to the standard deviation δ_s ($\Theta_{H'} = 2 \cdot \delta_s$, δ_s is defined in Eq. 3.7) so that the system has a high tolerance for the visual position difference when the quality of audio-motor maps is poor.

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

$$\begin{aligned} M^i_\theta(f,n,t) &= M^i_\theta(f,n,t-1) - F(\theta) \cdot \\ &\quad (M^i_\theta(f,n,t-1) - C'^i(f,n)), \end{aligned} \tag{3.10}$$

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

$$F(\theta) = c \cdot \beta \cdot \delta_{\theta,\theta_{v'}}, \tag{3.11}$$

where c and β represent the confidence of the matching 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 this work, β is set to 0.2 based on empirical tests, so that a good compromise is achieved. The position evidence vector $\delta_{\theta,\theta_{v'}}$ is defined by a delta function:

$$\delta_{\theta,\theta_{v'}} = \begin{cases} 1 & \text{if } \theta = \theta_{v'}, \\ 0 & \text{if } \theta \neq \theta_{v'}. \end{cases} \tag{3.12}$$

The online adaptation algorithm is described in Algorithm 4.

Algorithm 4 Online adaptation of audio-motor maps

Given population-coded cues $C'^{i}(f,n)$ in the current audio proto-object and position $\theta_{v'}$ in the matched visual proto-object
if $H < \Theta_H$ OR $H' < \Theta_{H'}$ **then**
    $c \leftarrow 1$ ▷ Calculate the confidence of an adaptation step
else
    $c \leftarrow 0$
end if
for $\theta = -90^\circ$ to 90° step 10° **do** ▷ Loop over azimuth angles
    if $\theta = \theta_{v'}$ **then**
        $\delta_{\theta,\theta_{v'}} \leftarrow 1$ ▷ Calculate the position evidence vector
    else
        $\delta_{\theta,\theta_{v'}} \leftarrow 0$
    end if
    Calculate learning parameter $F(\theta)$ based on Eq. 3.11
end for
for $\theta = -90^\circ$ to 90° step 10° **do** ▷ Loop over azimuth angles
    for $i = 1$ to 2 step 1 **do** ▷ Loop over cues (1 for IID, 2 for ITD)
        for $f = 1$ to 100 step 1 **do** ▷ Loop over frequency channels
            for $n = -0.9$ to 0.9 step 0.1 **do** ▷ Loop over node centers
                Update $M^i_\theta(f,n,t)$ based on Eq. 3.10
            end for
        end for
    end for
end for

3.4. Results

In this section, the performance of online adaptation of audio-motor maps using CAVSA is evaluated. In the investigated scenarios 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 FOV. Additionally, hardware modifications can lead to a change in the audio-motor maps. In order to test the adaptation of the system to modifications, the robot head was slightly modified and the exchange of left and right microphones was simulated. 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. For simplifying the description, the basic approach in Algorithm 2 is denoted as "*BASIC*" and the enhanced version in Algorithm 3 as "*IMPROVED*". These approaches were compared with a heuristic method denoted as "*HEU*" which 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. *HEU* behaves similarly to methods in [Hörnstein et al., 2006, Nakashima and Mukai, 2005] for linking auditory and visual signals.

3.4.1. Experimental Equipment

The system uses a humanoid robot head with a pair of cameras (Matrix Vision BlueFOX 224C) and a pair of microphones (DPA Miniature Microphones 4060-BM). The head is mounted on a pan-tilt unit. The cameras have a horizontal FOV of 90°. The microphones are located roughly at the position of human ears and surrounded by pinnae-like structures [Rodemann et al., 2008].

All experiments are conducted in a normal office room of size $4 \times 3 \times 2.8$ m with typical echo and background noise conditions such as noise from computers and air conditioner. As sound sources, pre-recorded utterances played from a loudspeaker or real human speakers are used. The loudspeaker is placed in front of the robot and at the same height as the robot head. Since this work does not rely on visual distance estimation, only the left camera is used. The distance between robot head and sound sources is about 1 m. This distance

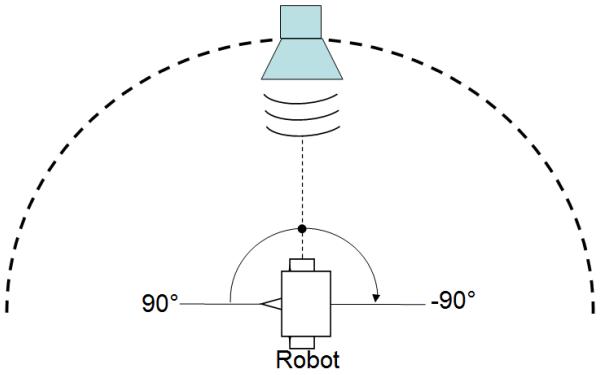


Figure 3.5.: Sketch of the experimental setting: offline calibration.

information and the saccade maps described in [Rodemann et al., 2006b] are used to convert camera coordinates to world coordinates.

3.4.2. 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°), as shown in Fig. 3.5. The robot head is rotated every 10° from $-90°$ to 90°. 47 sound files of average length 4.2 s are played at each grid position and the audio-motor maps are learned offline in a procedure outlined in [Yan et al., 2011b]. 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:

$$d(M,M') = \sqrt{\frac{\sum_{\theta,i,f,n}\left(M^i_\theta(f,n) - M'^i_\theta(f,n)\right)^2}{K}}, \tag{3.13}$$

where $M^i_\theta(f,n)$ and $M'^i_\theta(f,n)$ represent online-adapted and offline-calibrated maps respectively. K is the total number of elements in an audio-motor map and satisfies $K = k_\theta \cdot k_i \cdot k_f \cdot k_n$, where $k_\theta = 19$, $k_i = 2$, $k_f = 100$ 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

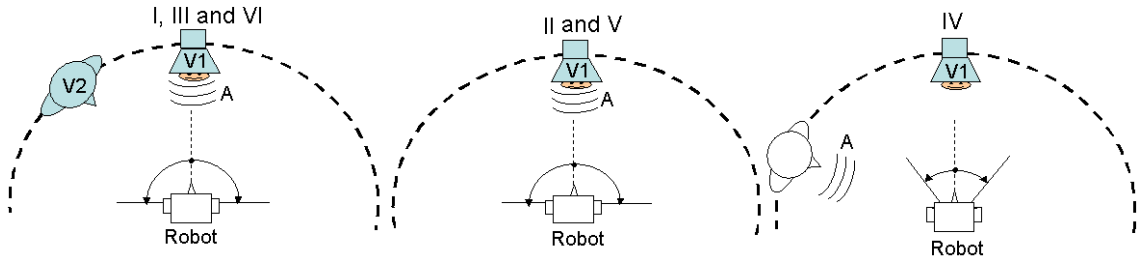


Figure 3.6.: Experiment 1: Sketch of the experimental setting in temporal phases I-VI. In phase IV, the range of the robot's head motion is $[-25°, 25°]$ and the current speaker has not yet been seen. In other phases, the head angle is in the range $[-90°, 90°]$ and the current speaker is the loudspeaker that has been seen. A is the current audio proto-object, V_1 and V_2 are visual proto-objects of the simulated participant and the additional person respectively.

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 main motivations for an online adaptation. In this case, offline-calibrated maps are not reliable to be used as reference, and the system also applies 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 position error E is described by:

$$E = \frac{\sum_{t=1}^{N} |\theta_t - \theta_{true}|}{N}, \tag{3.14}$$

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

3.4.3. Experiment 1: Consideration of Unseen Visual Proto-Objects

In the first experiment, the basic approach *BASIC* was compared with the heuristic method *HEU* and the approach *IMPROVED* that considers the situation where the matched visual proto-object is not in the STM.

The audio-motor map was initially filled with random numbers in the range $[-0.5, 0.5]$ using a uniform distribution. Fig. 3.6 sketches the experimental setting in temporal phases I-VI. Instead of a real person in the experiments, a loudspeaker was used on which a picture of

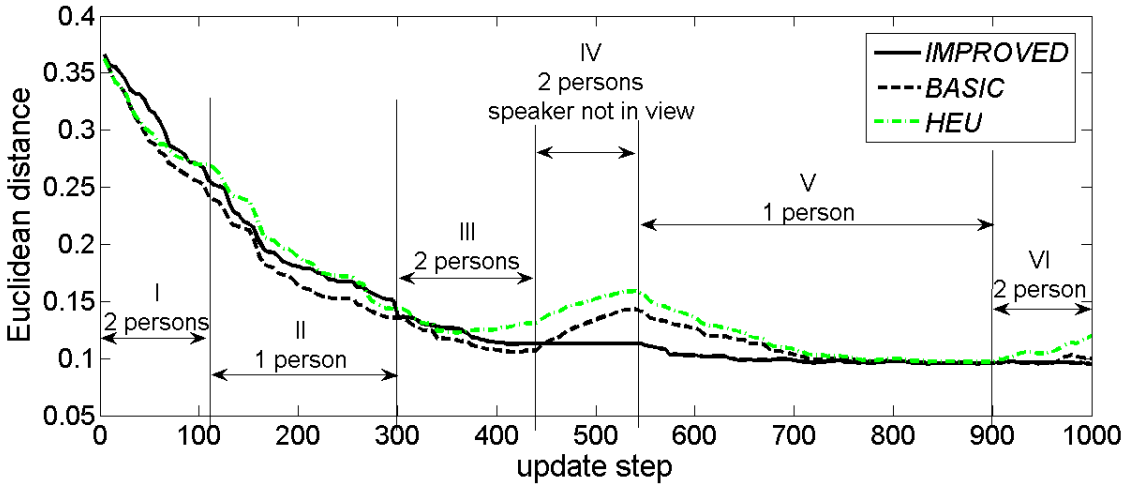


Figure 3.7.: Experiment 1: Comparison of three methods using Euclidean distance from offline-calibrated maps.

a face was attached, as the experiments were very time-consuming. The loudspeaker was placed in front of the robot ($0°$). During online adaptation, the robot head oriented itself to a random horizontal angle in the range $[-90°, 90°]$ after finishing each adaptation step, except that the head angle was in the range $[-25°, 25°]$ in phase IV. 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. An additional person dynamically entered and left the room. He talked to the robot only in phase IV, when the loudspeaker was turned off. Since the head angle was in the range $[-25°, 25°]$ during this time, the person as the sound source is not visible. In other phases, the person did not speak so that the only sound source was the loudspeaker at $0°$.

Fig. 3.7 shows a comparison of the three methods based on Euclidean distance from offline-calibrated maps in six phases. Fig. 3.8 illustrates the percentage of correctly learned (R), not learned (N) and wrongly learned (W) steps of different methods in each phase. Online adaptation with *HEU* was as good as that with *BASIC* and *IMPROVED* when only one visual proto-object was in the STM as in phase II and V, or when the quality of maps was still poor as in phase I. If more than one visual proto-object existed in the STM, *BASIC* and *IMPROVED* performed better than *HEU*, particularly when the maps were refined as in phase III and VI. If the matched visual proto-object was not in the STM as in phase IV, *HEU* and *BASIC* selected the wrong visual proto-object for audiovisual association, so that the quality of maps became poor. In comparison, *IMPROVED* refused audiovisual association and the performance of online adaptation did not decrease. The quality of online-adapted maps were

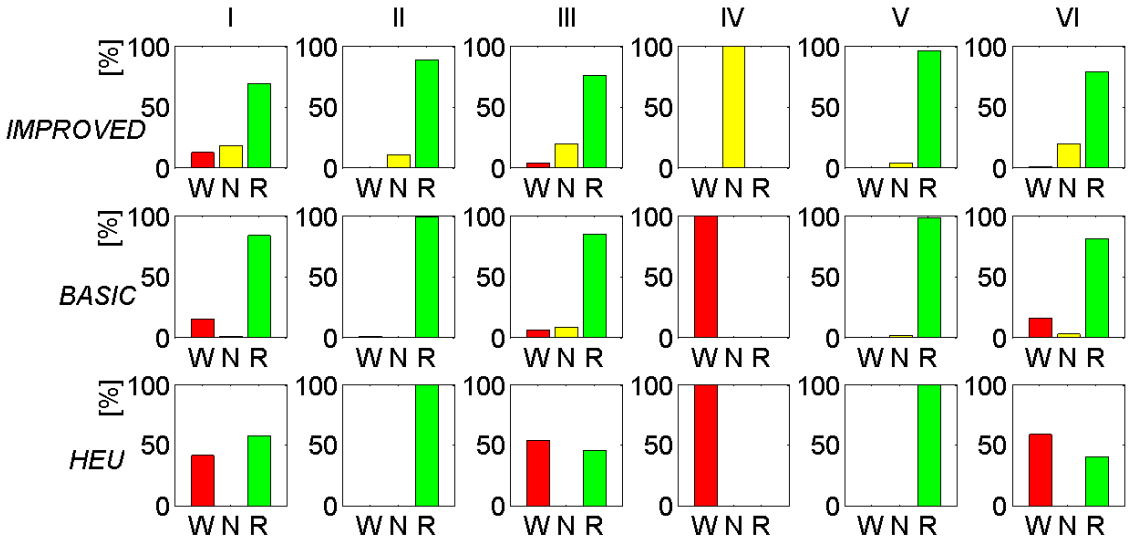


Figure 3.8.: Experiment 1: Percentage of update steps where a wrong visual proto-object (W), no visual proto-object (N) or the right visual proto-object (R) is chosen for audiovisual association in each phase.

also verified by using the maps to localize sounds from the calibration database. The average position error of 300 measurements using online-adapted maps with *IMPROVED* after 1000 update steps and offline-calibrated maps were 3.27° and 3.96° respectively. It is thus evident that online-adapted audio-motor maps performed as well as offline-calibrated maps, and that the results were valid for different performance metrics. In the following experiments, *IMPROVED* is applied .

3.4.4. Experiment 2: Bootstrapping of Online Adaptation

In the first experiment, it was noticed that the system is capable of bootstrapping in case of two persons. The ability of bootstrapping is further analyzed in a set of scenarios in the second experiment. The experimental setting is shown in Fig. 3.9. A loudspeaker was used on which a picture of a face was attached as in the first experiment. Up to 3 additional pictures of a face (V2, V3 and V4) were attached on objects that were also placed at a distance of 1 *m* 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 FOV and head movements. Since the only sound source was the loudspeaker, the ground truth position was 0°. Furthermore, the random azimuth angle of the robot head motion was in the range $[-90°, 90°]$ after finishing each adaptation step. As in the first experiment, the map was initialized with random numbers in the range $[-0.5, 0.5]$ using a uniform distribution. Therefore, it was not able to reliably map audio cues to source positions. Since there were

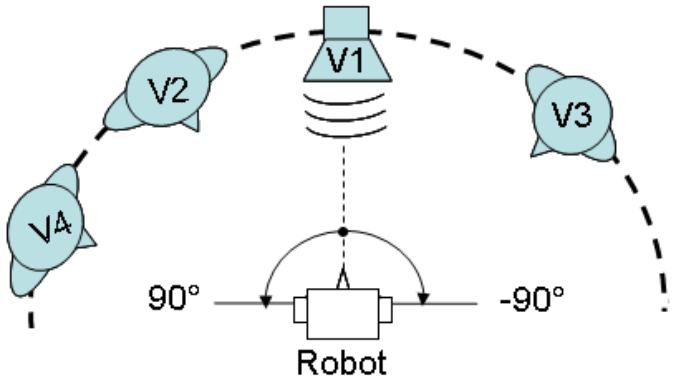


Figure 3.9.: Experiment 2: Sketch of the experimental setting for bootstrapping of online adaptation. V1 is the face picture attached on the loudspeaker which is the only sound source placed at 0°, while V2, V3 and V4 are additional face pictures. **Scenario I** contains only V1; **Scenario II** contains V1 and V2; **Scenario III**: V1, V2 and V3; **Scenario IV**: V1, V2, V3 and V4

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. This experiment was conducted to analyze this situation and to test the bootstrapping capability of the system. For these scenarios, the approach *IMPROVED* was compared with the heuristic method *HEU* which is similar to methods in [Hörnstein et al., 2006, Nakashima and Mukai, 2005] for linking auditory and visual signals. Also the method using known sound source position (0°) is shown as reference and denoted as "*IDEAL*".

Fig. 3.10 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. 3.11. We can see that *IMPROVED* was able to bootstrap itself in different scenarios and nearly reached the performance of *IDEAL*. *IMPROVED* outperformed *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. 3.11. In comparison, *IMPROVED* had only a small error rate mainly due to inaccurate sound localization. The small error rate of *IMPROVED* 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 association in this case and the person is treated as the current sound source. As shown in Fig. 3.10, the system almost reaches its final performance level (measured by Euclidean distance) after 400 steps, which took about 30 minutes, while offline calibration requires more than 2 hours.

The quality of online-adapted maps was also verified by using them to localize sounds from

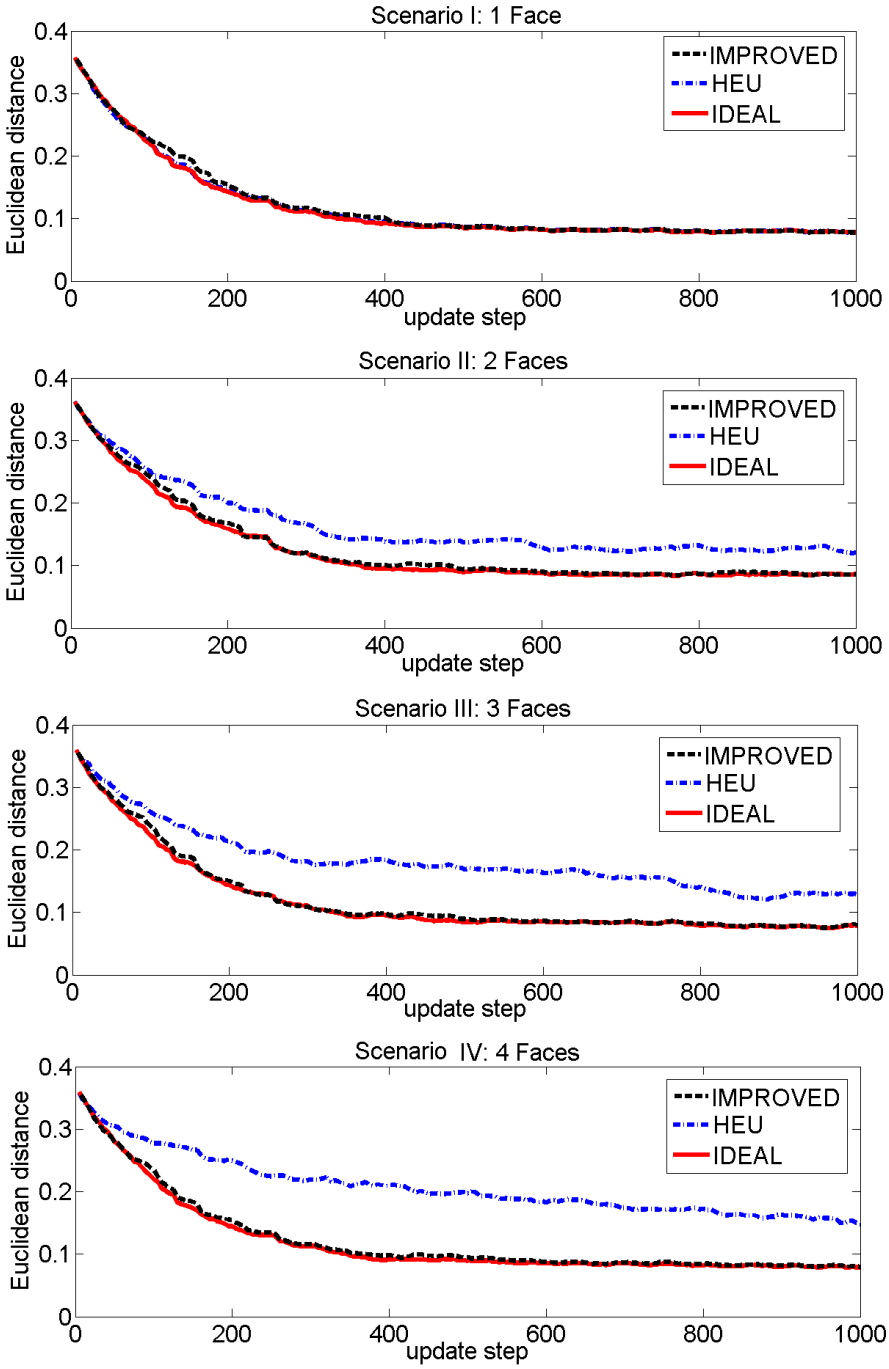


Figure 3.10.: Experiment 2: Comparison of bootstrapping in different scenarios with 1-4 faces using Euclidean distance to offline-calibrated maps

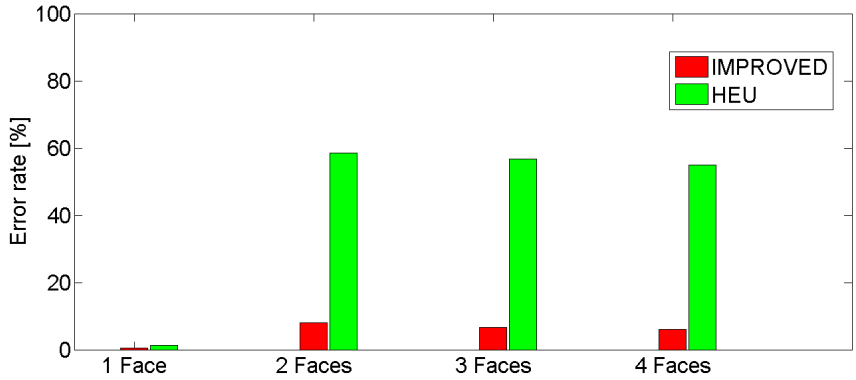


Figure 3.11.: Experiment 2: Percentage of update steps where a wrong "person" is chosen as the current speaker

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

Table 3.1.: Experiment 2: 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°.

a known position. The average absolute position errors after 1000 steps in different scenarios are shown in Table 3.1. The performance of offline maps is also given as reference. We can see that the results were consistent using different performance metrics.

Moreover, the learning of the audio-motor maps was tested 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 the system, because sound source separation is missing. When people talk at the same time, the sound position is measured as some average of all active speakers and the uncertainty of audiovisual association can be very high. In such confusing situations, entropy H can exceed its threshold Θ_H and audiovisual association is refused. Fig. 3.12 shows the Euclidean distance between online-adapted and offline-calibrated maps over time. It was 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 3.1. The errors after 10 and 1000 update steps are 61° and 22° , respectively. Fig. 3.13 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 to the offline-calibrated map. In comparison to scenarios as shown in Fig. 3.9, the learning progress was slow due to many complex and confusing situations. Therefore the maps need

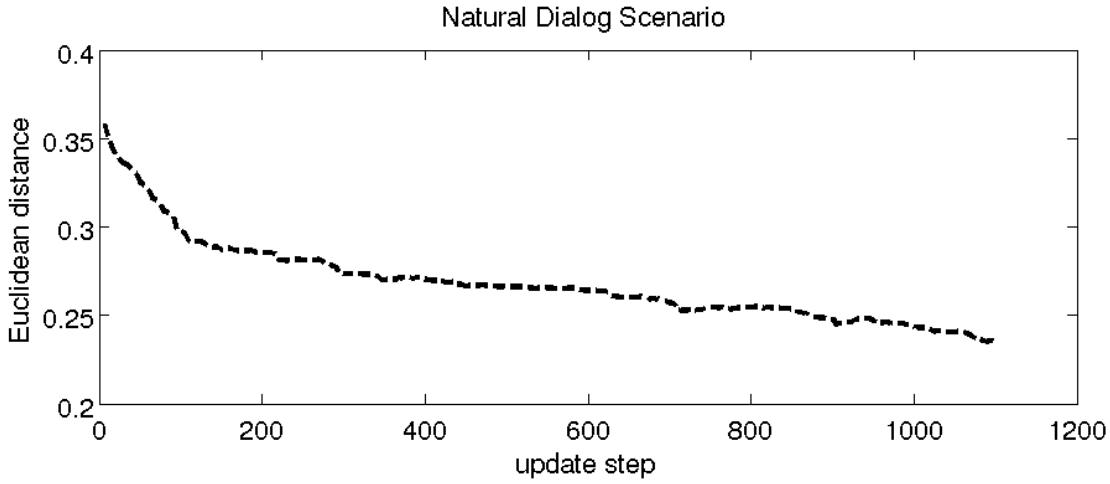


Figure 3.12.: Experiment 2: Euclidean distance between online-adapted and offline-calibrated maps in a natural dialogue scenario where four persons spoke freely without any script.

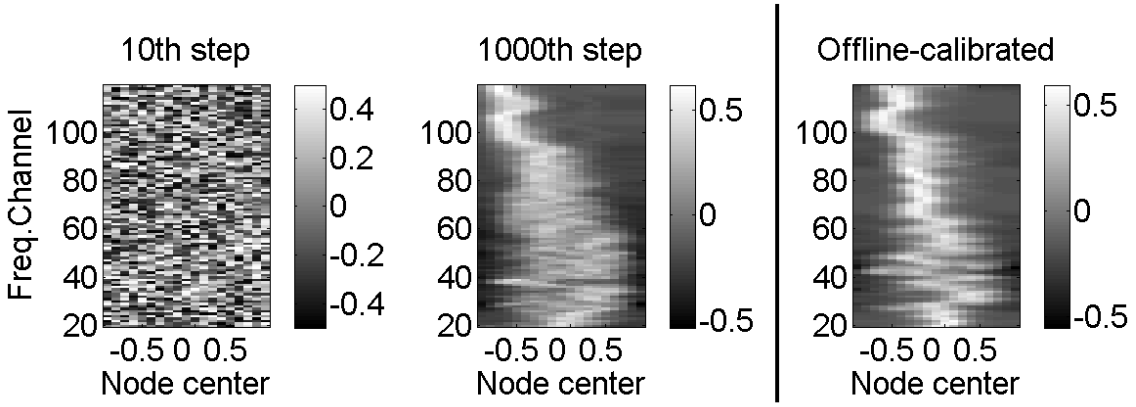


Figure 3.13.: Experiment 2: Online-adapted IID maps of $-20°$ after 10 and 1000 steps in the natural dialogue scenario with four speakers. Offline-calibrated maps are used as reference.

to be adapted longer than in relatively simple scenarios.

3.4.5. Experiment 3: Adaptation Ability

In the third experiment, it was 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. 3.9). 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 and the offline-calibrated audio-motor maps. For each update step, the mean position error in degrees was estimated using Eq. 3.14 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.

Modification of the Robot Head

The robot head was slightly modified by putting a paper hat over the outer ears to alter the mapping between audio cues and sound position. Fig. 3.14 shows the mean position error with offline-calibrated and online-adapted maps in scenario I with only one face. It is noticed 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 around 10°. Fig. 3.15 shows the comparison of sound localization performance between offline-calibrated and online-adapted maps in scenario II with two faces. In comparison to Fig. 3.14, 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%.

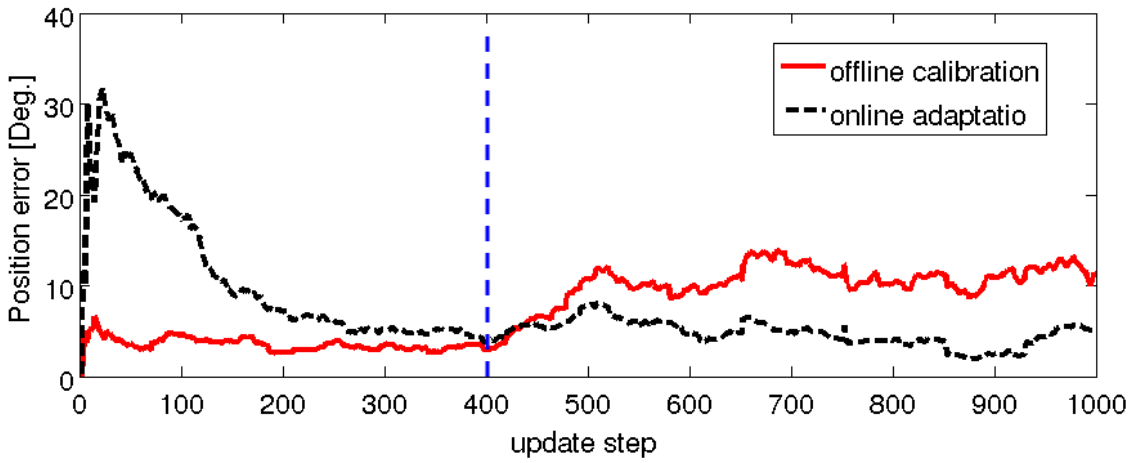


Figure 3.14.: Experiment 3: Mean position error with offline-calibrated and online-adapted maps in **scenario I**, when the robot head is modified in the 400th step.

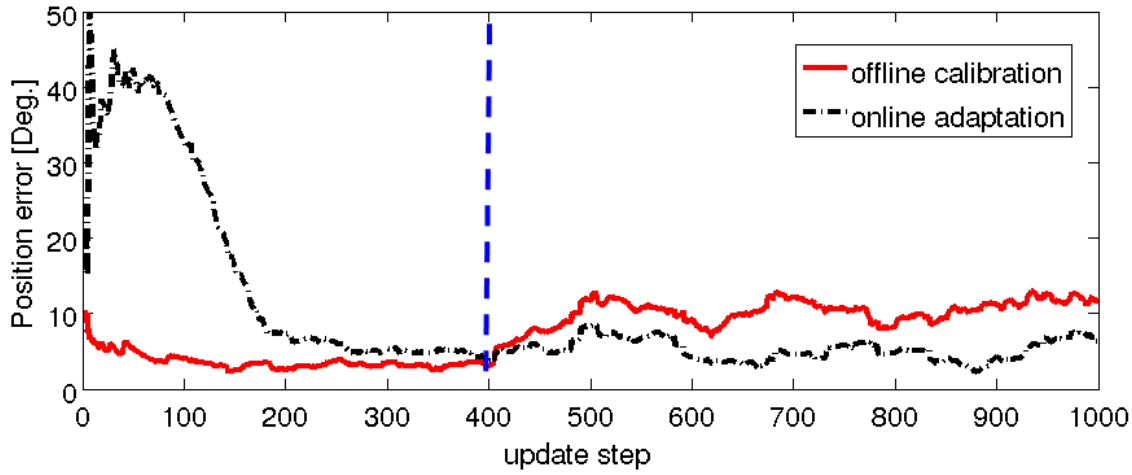


Figure 3.15.: Experiment 3: Mean position error with offline-calibrated and online-adapted maps in **scenario II**, when the robot head is modified in the 400th step .

Exchange of two Microphones

The exchange of left and right microphone was simulated 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. 3.16 (resp. Fig. 3.17) 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 is very time-consuming. 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. 3.14 and Fig. 3.15), while 2 hours have to be spent for repeating offline calibration. 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. 3.10), while only 50 minutes were spent for 1100 steps in a natural dialogue scenario (Fig. 3.12), because the participants talked all the time in the experiment and generated audio proto-objects faster. To summarize, it is beneficial to online adapt audio-motor maps instead of using offline calibration.

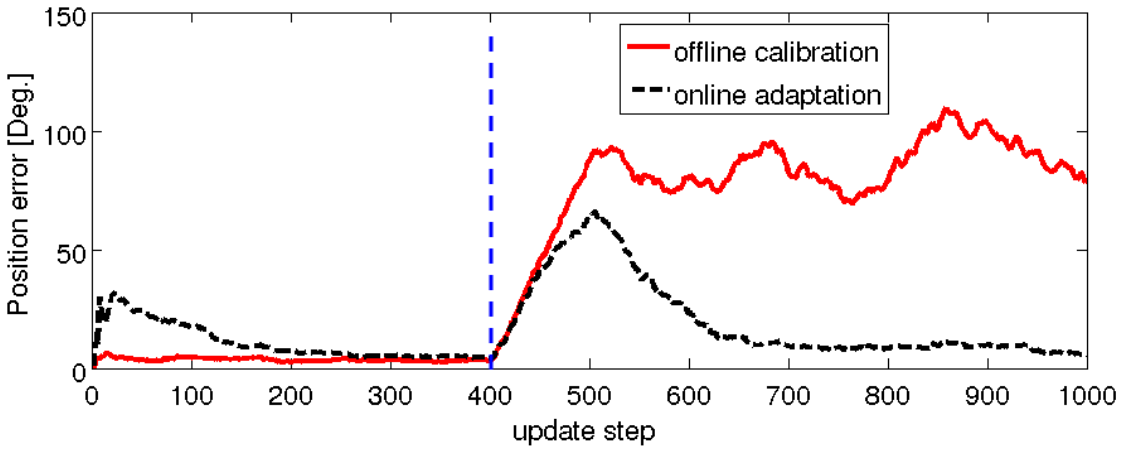


Figure 3.16.: Experiment 3: Mean position error with offline-calibrated and online-adapted maps in **scenario I**, when the exchange of two microphones is simulated in the 400th step.

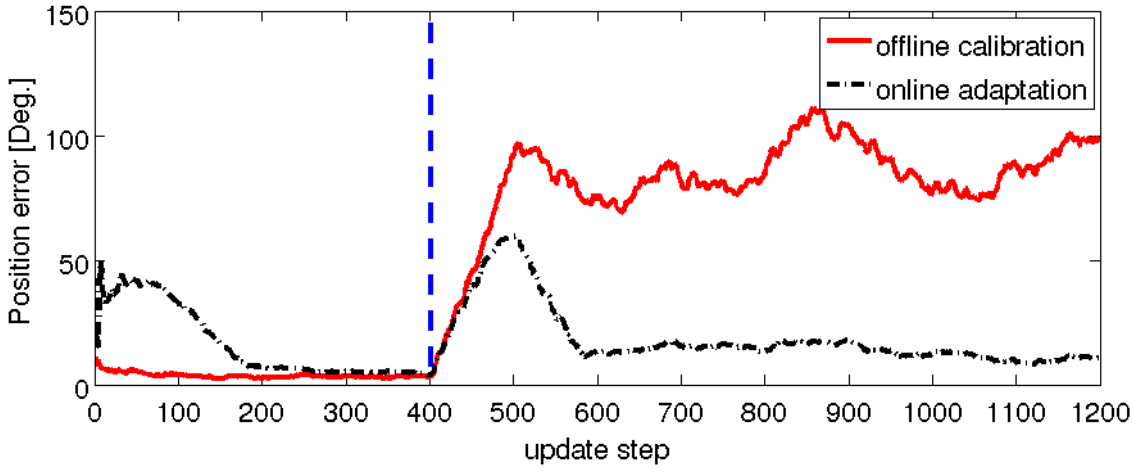


Figure 3.17.: Experiment 3: Mean position error with offline-calibrated and online-adapted maps in **scenario II**, when the exchange of two microphones is simulated in the 400th step.

3.5. Discussion

CAVSA was applied to search for the matched visual source to the current sound in multi-person environments. Comparing the proposed system with the state of the art in terms of learning progress, it has been shown that the approach with CAVSA was more robust in multi-person scenarios. It has also been shown that the performance of the system with *IMPROVED* does not degrade when the related visual source has not yet been spotted. Furthermore, the proposed 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, the approach with CAVSA was capable of bootstrapping itself in a natural dialogue scenario with 4 human speakers. The learning progress was slow but steady in this complex scenario. In case of hardware modifications, the adaptation process was capable of quickly restoring the old localization performance in the single-person scenario. In the multi-person scenario, the adaptation process nearly reached the same performance, when the mapping was slightly modified by putting a hat on the robot head. When the mapping was changed so drastically that left and right 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. Thanks to the adaptation ability in case of the modifications of acoustic conditions, we do not need to spend 2 hours for repeating the whole process of offline calibration. This is one main motivation to online adapt audio-motor maps.

The system does not need control over the robot motion and thus can work during other tasks. Nevertheless, 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). It is assumed that most relative azimuth angles are generated through the robot's head motion and movements of the interacting partners during various tasks. In the 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 [Karaoguz et al., 2013]. For example the proposed 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 association 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.

The current version of CAVSA works well under the assumption that people do not change their position much in a dialogue scenario. If a visual target disappears for a while and then reappears from a different place or moves quickly, it will be considered as a new one. This is because only position information is used to group visual proto-objects in the STM. Although this does not affect the performance of online adapted audio-motor maps, the scene representation is erroneous. Therefore, more grouping features such as color and texture are employed in dialogue scenarios in the next chapter.

4. CAVSA in Human-Robot Dialogue Scenarios

In many real-world situations, a robot is interacting with multiple people. In this case, understanding of dialogues is essential. However, dialogue scene analysis is missing in most existing systems of human-robot interaction. In such systems, only one speaker can talk with the robot or each speaker has to wear an attached microphone or a headset. Therefore, CAVSA is applied to make dialogues between humans and robots more natural and flexible. It aims at learning the number and position of interacting partners, as well as who is currently speaking. CAVSA is a challenging task as described in section 1.1.

In the previous chapter, the original version of CAVSA was tested in online adaptation of audio-motor maps. Only a small set of simple features was sufficient since there was no need to estimate person IDs. In dialogue scenarios, although it may not be important to figure out exactly who a person is, the system has to notice when the camera sees the same person again after he left the scenario for a while. In the example given in section 1.1 a robot is working behind a bar. When a guest has ordered a drink, leaves the bar shortly and returns, the robot should recognize him again. Otherwise, the robot could give a drink to the wrong guest. However, the original version of CAVSA fails in this situation since only position information is used to group visual proto-objects in the STM. Therefore, more visual features are employed to recognize person identities for improving the capabilities of CAVSA. Similarly to visual features, a set of simple auditory features is tested to recognize voices. Furthermore, auditory features are applied to enhance the system robustness in terms of filtering out typical environmental sounds such as sounds of phone bells, door creaking (when opening or closing the door) and placing a cup on the table.

The outline of this chapter is as follows: section 4.1 gives an overview of related work in audiovisual human-robot interaction with multiple persons. Then the extraction of visual features on different upper body parts, feature selection and feature fusion are introduced in

section 4.2. Section 4.3 explains the extraction and selection of auditory features. After that, CAVSA is set adapted to dialogue scenarios in section 4.4, and section 4.5 shows results in several dialogue scenarios. Finally, section 4.6 will cap this chapter with a discussion.

Important aspects of combining more visual features and the corresponding results were already published in [Yan et al., 2012].

4.1. Related Work in Audiovisual Human-Robot Interaction with Multiple Persons

Much research has been carried out around the area of audiovisual human-robot interaction with multiple persons, such as audiovisual association based speaker detection [Haider and Moubayed, 2012, Sanchez-Riera et al., 2012], tracking of speakers [Kim et al., 2007, Checka et al., 2003, Khalidov et al., 2008, Nakadai et al., 2001, Fritsch et al., 2003] and audiovisual attention control [Bennewitz et al., 2005, Yan et al., 2013b]. These methods can also enable a robot to determine the number and position of persons as well as who is the current speaker, but work in constrained scenarios. [Haider and Moubayed, 2012] can detect the current speaker based on synchrony between speech and lip movements. However, this method works only when speakers are in the camera field of view (FOV). Since most robots have a limited camera FOV, a speaker may stand out of view for a while due to speaker's or robot's movement. Even when a speaker is in the camera FOV, he could be occluded for a moment. In those cases, most of these methods such as [Sanchez-Riera et al., 2012, Kim et al., 2007, Checka et al., 2003, Khalidov et al., 2008] do not memorize the speaker and would consider him as a new person. Although [Nakadai et al., 2001] can recognize persons using face recognition, the training of faces is required beforehand. Please notice that the STM described in [Nakadai et al., 2001], unlike the STM used in this work, is just a cache that stores all recently received signals. [Bennewitz et al., 2005] proposed an approach introducing a probabilistic belief about people in the surroundings of the robot. While the robot head is moving, a person may leave the FOV briefly, but his face is still stored in the belief and the robot can find him again later. However, since the observation assignment to stored faces is purely based on their positions, the method fails if a person greatly alters his position outside the FOV. [Fritsch et al., 2003] used multi-model anchoring to assign observations to symbolic objects. New anchoring processes are established if observations cannot be assigned to the existing symbolic objects. Contrary, anchoring processes are removed if no observations are assigned for a certain period of time. The memorization mechanism is

similar to the STM used in this work. In addition, [Fritsch et al., 2003] can identify faces with pre-training as [Nakadai et al., 2001]. However, when people are unknown in advance, the system assigns observations to objects (speakers) based on only position information. Hence, it has the same drawback as [Bennewitz et al., 2005] in this case. [Yan et al., 2013b] applied a visual STM and used also only position information to group persons in the STM. Moreover, [Yan et al., 2013b] is proposed for meeting scenarios which are simpler than typical dialogue scenarios. Please note that Yan from [Yan et al., 2013b] is different from the author of this thesis. In their meeting scenarios, the number of participants is constant, people have fixed seats and therefore do not change their position much. The robot only has to initially turn around once to store the positions of all persons in the meeting room into the STM. In comparison to the STM described in this work, the "insert" and "remove" step are not required. To summarize, the approaches [Bennewitz et al., 2005, Fritsch et al., 2003, Yan et al., 2013b], in general, are more similar to the original version of CAVSA, given people unknown to the system beforehand.

4.2. Visual Features

Additional visual features are required to differentiate multiple people in dialogue scenarios. Since the system deals with an unknown number of persons who have not been seen and heard before, a database for offline training of high-level features to recognize faces is not available. Furthermore, high-level features for face recognition exhibit an inherent computational complexity which makes real-time processing difficult. Hence the system combines a set of simple visual features such as height, color and texture of different upper body parts. Multiple features of the same person can better represent the person and increase the recognition accuracy because of redundant information. When an individual feature is ambiguous to differentiate multiple people, the robustness is enhanced by their combination.

The presence of persons is detected by searching for their faces. Based on face position and size, areas of hair, collar and clothes are estimated (section 4.2.1). In these areas color and texture features are extracted in form of histogram vectors (section 4.2.2). To keep the same metric for all visual features, visual camera positions are converted to position evidence vectors in spherical coordinates (section 4.2.3). In this manner, all features with the same structure can easily be processed. Furthermore, the position evidence vector of the elevation angle is used as the height feature. Finally, a feature subset is selected to achieve the best recognition performance, and a fusion operator is chosen (section 4.2.4).

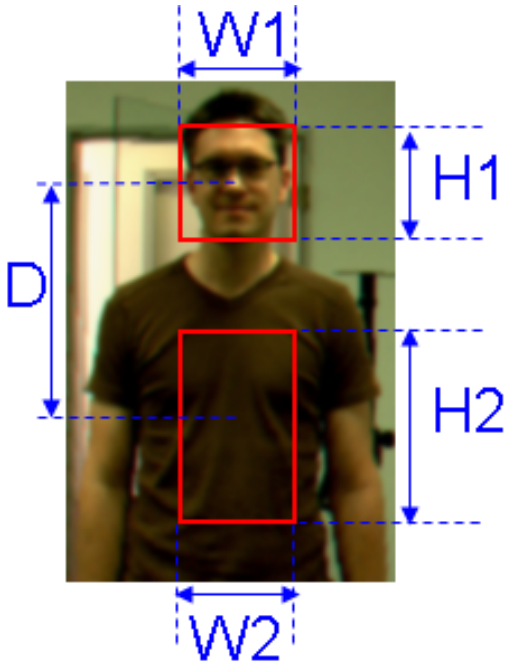


Figure 4.1.: Description of the parameters used for the estimation of upper body parts (e.g. clothes ROI), where visual features will be extracted.

4.2.1. Face-based Detection of Different Upper Body Parts

People usually do not change clothes or hairstyle during a dialogue scenario. Hence people can be distinguished using color and texture features of different upper body parts (hair, face, collar and clothes). The border region between the face and hair region is more useful for differentiating between people than the bulk texture and color of a person's hair on its own [Corcoran et al., 2005]. Hence the selected hair area should contain most of the border region and only little background. Similarly, a collar region is extracted as the border between neck and clothes which contains much information about the clothing style. For instance, a shirt, coat and blouse can be distinguished only by the collar. The area below the collar is considered as clothes area.

In the camera FOV, the presence of persons is detected by searching for their faces as described in section 3.2.2. For each face, the center and size in the camera image is computed. The upper body parts are extracted according to the position and size of the detected faces. The width and height of each region of interest (ROI) are proportional to the one of the corresponding face ROI, and the distance to the face ROI is proportional to face height. In this manner, the scale of different upper body parts on the camera image changes coherently as the distance or viewing angle is changing.

Figure 4.2.: Examples: face-based detection of clothes ROI

Fig. 4.1 shows an example of parameters for determining the clothes ROI. Three ratio parameters $r_W = \frac{W_2}{W_1}$, $r_H = \frac{H_2}{H_1}$ and $r_D = \frac{D}{H_1}$ are applied, where W_1, H_1, W_2, H_2 and D are the width and height of the face ROI, width and height of the clothes ROI, and the distance between the face and clothes ROI, respectively. Using the criterion that most of the desired content and only little background appear in the ROI, these three parameters are chosen by training on several images of ten subjects (see Fig. 4.2). r_W, r_H and r_D for each ROI are shown in Table 4.1 (see also [Yan et al., 2012]). Fig. 4.2, Fig. 4.3 and Fig. 4.4 illustrate examples of hair, collar and clothes ROIs respectively which are estimated using these parameters. Similar methods of face-based clothes detection can be found in [Jaffre et al., 2004, Fritsch et al., 2004].

	Hair	Collar	Clothes
r_W	1.2	1.2	1.2
r_H	0.3	0.8	1.67
r_D	0.5	1	2.1

Table 4.1.: Parameters for estimation of hair, collar and clothes ROI

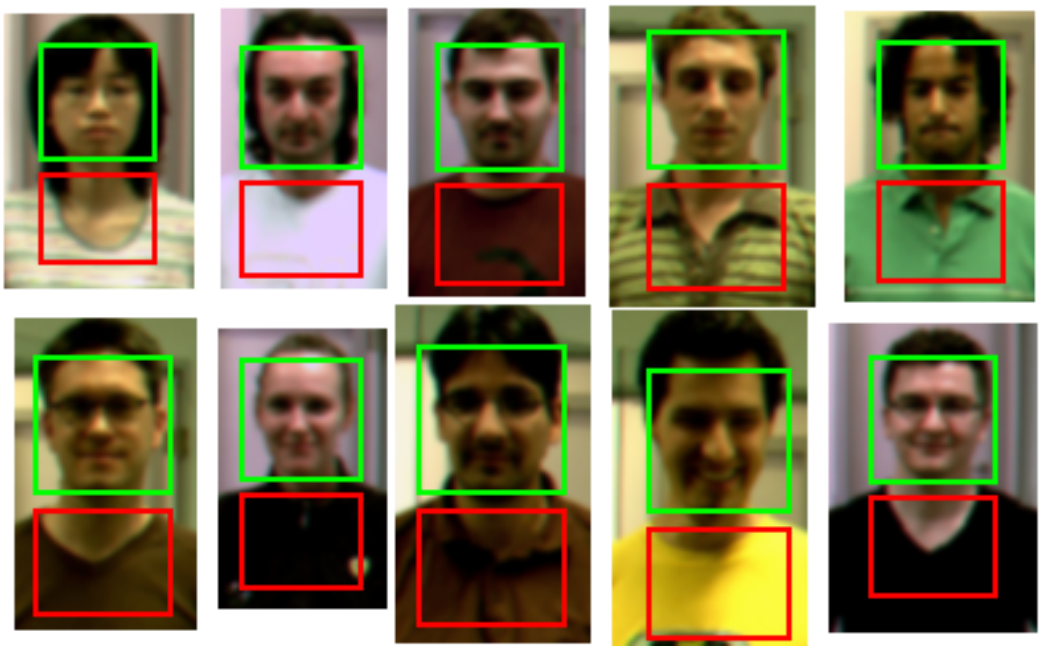


Figure 4.3.: Examples: face-based detection of collar ROI

Figure 4.4.: Examples: face-based detection of hair ROI

4.2.2. Histogram Feature Vector of Color and Texture

Histograms are among the most common features used in various tasks such as pattern recognition, tracking and segmentation. From a probabilistic point of view, a normalized histogram is similar to a probability density function (PDF), which is one of the most popular pattern representations [Cha, 2007]. A histogram shows basic information about a data set, such as the kind of distribution (the histogram shape), statistical properties of the distribution (e.g. width of spread) and outliers in the data set. Moreover, histograms are normalized by the number of pixels in the ROI, so that histograms represent the ROI without regard to its scale. The system calculates color histograms in RGB color space as well as Local Binary Pattern (LBP) histograms for the above mentioned upper body parts. To simplify the description, these histogram features are denoted as Hair RGB, Hair LBP, Face RGB, Face LBP, Collar RGB, Collar LBP, Clothes RGB and Clothes LBP respectively in the following text.

Color Histogram

Color histograms have been proven to be a robust and efficient cue for object representations [Swain and Ballard, 1991]. They are stable in the presence of occlusion and over change of viewing angles. The original form of color histograms are obtained by discretizing the image colors and counting the number of pixels that have colors in each of a fixed list of color ranges [Swain and Ballard, 1991]. The drawback of using such a color histogram is that a change in lighting conditions may lead to a corresponding shift in the histogram and thus histogram similarities are completely misjudged [Schettini et al., 2001]. As a perfect individual feature is not the key point of this work and the lighting condition of the experiment room is stable, the original form of color histograms is used for simplicity. Cumulative histograms could be used to make the representation robust to lighting changes [Stricker and Orengo, 1995]. The color histogram is calculated in RGB color space. Although RGB color space is neither perceptually uniform nor intuitive, it provides a simple and fast computation [Park et al., 1999]. Color histograms calculated in other color spaces can be found e.g. in [Park et al., 1999]. In this work, each color channel is discretized into six bins and there are a total of 216 (6^3) bins. The 3D histogram is concatenated into a single vector with 216 elements, because the computation of the distance/similarity between multi-dimensional histograms is very expensive.

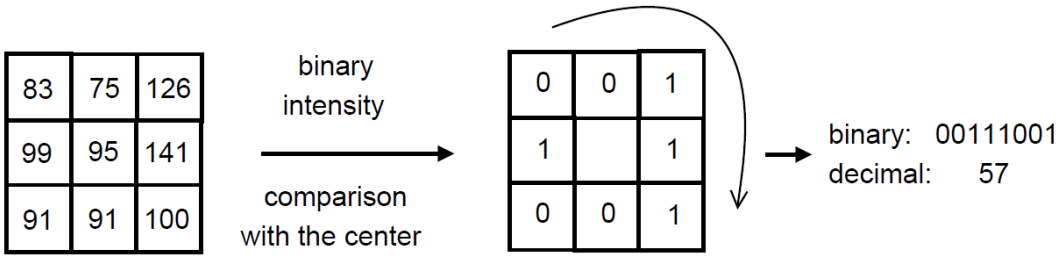


Figure 4.5.: Calculation of the LBP [Marcel et al., 2006]

Local Binary Pattern Histogram

The Local Binary Pattern (LBP) proposed by Ojala et al. is a powerful feature for texture classification [Ojala et al., 1996]. The CAVSA system uses LBP to represent local structure for its simplicity and robustness to luminance variation. As shown in Fig. 4.5, LBP is defined as an ordered set of binary comparisons of pixel intensities between a pixel and its eight surrounding pixels. Given i_c as the grey value of the center pixel and i_n ($n \in [0,7]$) as the grey values of eight surrounding pixels, the decimal form of the LBP code is given by [Marcel et al., 2006]:

$$LBP = \sum_{n=0}^{n=7} s(i_n - i_c) \cdot 2^n, \tag{4.1}$$

where function $s(i_n - i_c)$ converts the differences of grey values $i_n - i_c$ to binary numbers:

$$s(i_n - i_c) = \begin{cases} 1 & \text{if } i_n - i_c \geq 0, \\ 0 & \text{if } i_n - i_c < 0. \end{cases} \tag{4.2}$$

The system uses a subset of LBP called uniform patterns which represent most texture information such as corner, line and edge [Topi et al., 2000]. There are only 58 LBP patterns in the subset while the whole LBP set has 256 (2^8) elements. The subset contains at most two bitwise 0 to 1 or 1 to 0 transitions [Topi et al., 2000]. For example, uniform patterns 00000000, 00011111 and 00110000 have 0, 1 and 2 transition(s), respectively. The number of histogram bins is set to 58 (the number of elements in the subset).

4.2.3. Position Evidence Vector in Spherical Coordinates

Visual camera positions need to be converted to position evidence vectors in spherical coordinates, so that the same metric is kept for all visual features. The same format of features

simplifies the system structure and eases processing. First, camera positions are converted into 3D world coordinates and spherical coordinates (azimuth and elevation) as described in section 3.2.2. The estimated azimuth angle θ_v is then converted to a position evidence vector $p_v(\theta)$ for each angle θ (-90, -85, ..., 0, ..., 85, 90). As the visual position has higher precision than the auditory one, visual angles are more densely sampled. $p_v(\theta)$ is defined by a Gaussian function:

$$p_v(\theta) = \exp\left(\frac{-(\theta-\theta_v)^2}{2\cdot\sigma_\theta^2}\right), \tag{4.3}$$

where σ_θ is measured empirically to tolerate position measurement errors of the face detector and conversion errors from camera coordinates to world coordinates. Furthermore, since position is one among many features to group visual proto-objects in STM, slight head movements should also be tolerated. In consideration of all above aspects, σ_θ is set to 5 in the experiments.

Similarly, elevation angles are also converted to position evidence vectors for each elevation angle θ' $(-30,-25,...,0,...,25,30)$ and used as height feature. The deviation $\sigma_{\theta'}$ is set to 3 under consideration that the vertical head movements are normally smaller than horizontal ones.

4.2.4. Feature Subset Selection

The extraction of visual features (height, color and texture of different upper body parts) was introduced in the previous sections. When all these features are applied, the computational cost is high and recognition performance is not necessarily better than for only a subset of them. Therefore a feature subset is selected to achieve the best performance. The selection is carried out based on the database: 30 frames for each of ten subjects (see Fig. 4.2), 10 for training of the template and 20 as samples for testing.

Let D be the original set of features, $X_k = x_1, x_2, ..., x_k$ the feature subset which contains k features and $J(X_k)$ the feature selection criterion function. If a higher value of J indicates a better feature subset, the problem of feature subset selection is to find $X_k \subseteq D$ such that $J(X_k)$ is maximum.

Algorithm 5 SFFS algorithm

1: initialize $k = 0$
2: Search for feature x_{k+1} that satisfies $x_{k+1} = \underset{x_i \in D \setminus X_k}{\arg\max}(J(X_k \cup x_i))$
3: $X_{k+1} := X_k \cup x_{k+1}$
4: **if** k<2 **then**
5: $k \leftarrow k+1$
6: go to 2
7: **else**
8: Search for feature x_j that satisfies $x_j = \underset{x_i \in X_{k+1}}{\arg\max}(J(X_{k+1} \setminus x_i))$
9: **if** $x_j = x_{k+1}$ **then**
10: $k \leftarrow k+1$
11: go to 2
12: **else**
13: $X'_k := X_{k+1} \setminus x_j$
14: **if** $J(X'_k) > J(X_k)$ **then**
15: $X_k \leftarrow X'_k$
16: Search for feature x_r that satisfies $x_r = \underset{x_i \in X_{k+1}}{\arg\max}(J(X_k \setminus x_i))$
17: $X'_{k-1} := X_k \setminus x_r$
18: **if** $J(X'_{k-1}) > J(X_{k-1})$ **then**
19: $X_{k-1} \leftarrow X'_{k-1}$
20: $k \leftarrow k-1$
21: **if** $k = 2$ **then**
22: go to 2
23: **else**
24: go to 16
25:
26: **end if**
27: **end if**
28: **else**
29: go to 2
30: **end if**
31: **end if**
32: **end if**

Related Work in Feature Subset Selection

There are a large number of algorithms proposed for performing feature subset selection, for a survey see [Jain and Zongker, 1997]. An exhaustive approach can guarantee the optimal subset, but would require examining all possible subsets (2^n). The exponential search is impractical even for a moderate sized D. Alternatively, the Branch And Bound algorithm can be used to find the optimal feature subset much more quickly [Jain and Zongker, 1997]. For this, it requires the function J to be monotonic. This means that for any two subsets X_A, X_B and $X_A \subseteq X_B$, $J(X_A) \leq J(X_B)$ has to hold. However, in our case it might actually not always be true that the addition of new features increases the value of J. Additionally, the most commonly used methods are sequential feature selection algorithms because of their low cost. The sequential forward/backward selection algorithm starts with the empty/full set and adds/removes one feature in each step to make J bigger. These methods have comparable performance but do not guarantee an optimal result [Jain and Zongker, 1997]. This is because for example once a sequential forward selection (SFS) operator selects a feature, it will never be able to remove it again. Depending on the order of features it might therefore easily miss the optimal subset. In contrast, the sequential forward floating selection (SFFS) algorithm adds features based on sequential forward selection and allows to remove the worst features in the newly updated set [Pudil et al., 1994]. According to [Jain and Zongker, 1997], the performance of the floating method is comparable to the Branch And Bound algorithm and its speed is faster in most cases. Therefore, the SFFS algorithm is applied to select features, as described in Algorithm 5.

Feature Selection Criterion Function

The SFFS algorithm is used to select features. First of all, the feature selection criterion function J needs to be defined. For this purpose, we calculate the TPR (true positive rate), FPR (false positive rate) and analyse the ROC (receiver operating characteristic) for each examined subset. ROC is a metric for comparing predicted and actual target values in a binary classification model. If plotted, each point represents a pair of the TPR (on the y axis) and FPR (on the x axis) for a particular discrimination threshold. The TPR defines the number of correct positive results in proportion to all positive samples available during the classification test. The FPR, on the other hand, defines the number of incorrect positive results in proportion to all negative samples available during the test. There are many criteria to evaluate ROC points. Here, the point which has the minimum distance to the perfect

classification (TPR= 1 and FPR= 0) is selected and the minimum distance is denoted as d_m. The feature selection criterion function is defined as $J = 1 - d_m$, since we are maximizing this function. Additionally, the discrimination threshold Θ_s is set from 0 to 1 in steps of 0.001. Calculation of TPR and FPR for the large set of thresholds is time consuming, but visual feature selection is done only once in an offline manner. Hence the speed of the feature selection is much less important than the classification performance of the selected feature subset.

Feature Similarity and Fusion Operator

In the ROC-analysis, all features of one subject, namely Hair RGB, Hair LBP, Face RGB, Face LBP, Collar RGB, Collar LBP, Clothes RGB, Clothes LBP and height feature, build a visual proto-object. Whether a sample belongs to a template is decided by comparing the similarity of their proto-objects with a threshold Θ_s. Different metrics can be utilized to calculate vector similarity, for a survey see [Cha, 2007]. Given F as the number of features in a proto-object, the normalized inner product (NIP) is used to compute **feature similarity** s_i ($i \in [1,F]$) :

$$s_i = \frac{x_i \cdot y_i}{||x_i||\,||y_i||}, \tag{4.4}$$

where x_i and y_i stand for the i-th feature vector in a sample and in a template, respectively. The classification performance of each individual feature (set $F = 1$) is evaluated by d_m in Table 4.2. Features are arranged according to d_m in the table and the corresponding TPR, FPR are shown.

Feature	d_m	TPR	FPR
Hair LBP	0.1892	0.8693	0.1368
Height	0.1995	0.8693	0.1508
Clothes LBP	0.2204	0.8643	0.1736
Clothes RGB	0.2218	0.8844	0.1893
Collar LBP	0.2265	0.8894	0.1977
Collar RGB	0.2878	0.8241	0.2278
Face LBP	0.2955	0.7789	0.1960
Face RGB	0.3971	0.6382	0.1636
Hair RGB	0.4617	0.6080	0.2440

Table 4.2.: The minimum distance d_m to the perfect classification on the ROC-curve and the corresponding TPR, FPR for each individual feature

When more than one feature is used ($F > 1$), the measurement of proto-object similarity S

Step	Added Feature	d_m	TPR	NPR	Threshold
1	**Hair LBP**	0.1892	0.8693	0.1368	0.8020
2	**Height**	0.0814	0.9447	0.0597	0.8710
3	**Clothes LBP**	0.0463	0.9698	0.0352	0.8560
4	**Face LBP**	**0.0329**	**0.9749**	**0.0212**	**0.8550**
5	Collar LBP	0.0358	0.9799	0.0296	0.8410
6	Face RGB	0.0451	0.9749	0.0374	0.8200
7	Hair RGB	0.0623	0.9397	0.0156	0.8320
8	Clothes RGB	0.1358	0.8643	0.0045	0.8220
9	Collar RGB	0.1709	0.8291	0.0039	0.8220

Table 4.3.: SFFS algorithm for feature subset selection

between a sample and a template is involving the fusion of feature similarity s_i. The non-trainable **fusion operator** could be for example the mean operator:

$$S = \frac{1}{F} \cdot \sum_{i=1}^{F} s_i, \tag{4.5}$$

or the product operator:

$$S = \prod_{i=1}^{F} s_i. \tag{4.6}$$

First, the mean operator and the SFFS algorithm are used to select the feature subset. The feature selected in each step is shown in Table 4.3. We can see that the best feature subset contains Hair LBP, Face LBP, Clothes LBP and height feature. No features are excluded in the feature selection process, so that the result of SFFS is not different from SFS. We can see that Clothes RGB is not in the best subset, although it has better performance than Face LBP when using individual features. A possible reason is that Clothes RGB provides redundant information for the subsets with Hair LBP, Clothes LBP and height feature, based on our database. Actually, all features should be uncorrelated except for Clothes RGB and Collar RGB. Fig. 4.6 illustrates the ROC-curves for Hair LBP, Face LBP, Clothes LBP, height and their combination. According to [Zweig and Campbell, 1993], the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the classification test. As can be seen, the curve of the combined features is much better than that of any individual features.

We also run the SFFS algorithm using the product operator. The best feature subset is the combination of Hair LBP, Clothes LBP and height. The minimum distance d_m to the best classification is 0.0447, which is larger than $d_m = 0.0329$ for the best subset using the mean operator. Therefore, the mean operator is chosen for feature fusion, and the best feature subset (Hair LBP, Face LBP, Clothes LBP, height) is used for visual proto-objects.

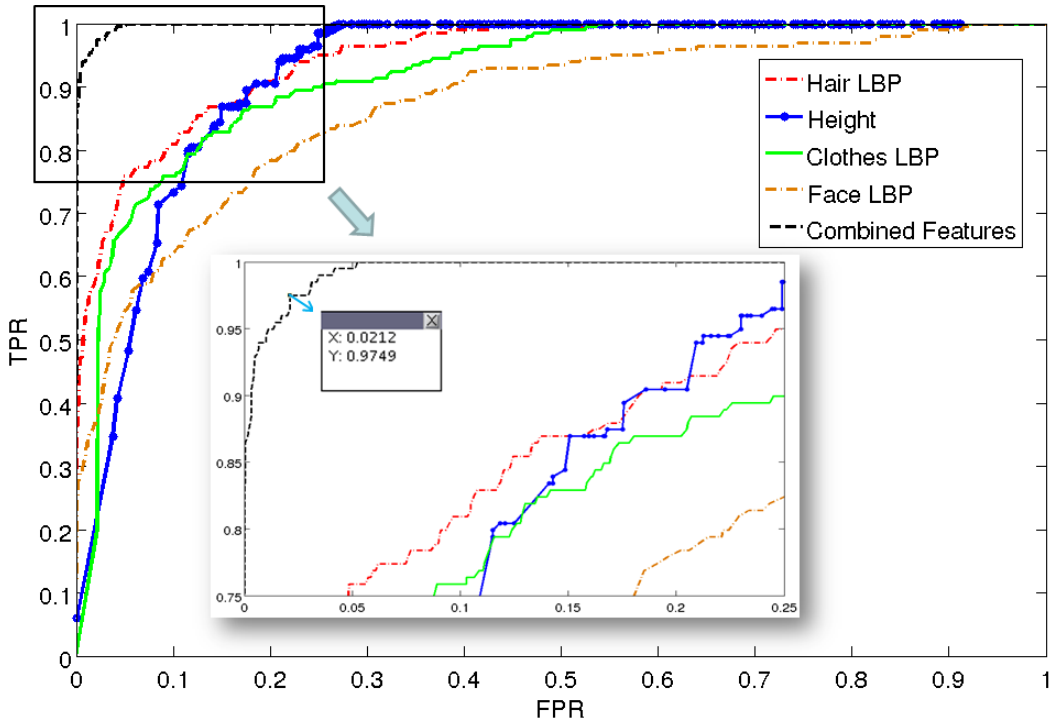


Figure 4.6.: ROC-curves for Hair LBP, Face LBP, Clothes LBP, height and their combination. The mean operator is used for the feature fusion. The point $(0.0212, 0.9749)$ on the curve of combined features has the minimum distance d_m to the perfect classification, $d_m = 0.0329$.

4.3. Auditory Features

The purpose of using auditory features is to improve the system robustness in terms of voice recognition and filtering out environmental sounds. When a speaker leaves the scene and continuously talks to the robot from outside its FOV, the system should notice that he has been heard before. Therefore, several simple auditory features are tested to identify voices, similar to what was done for the visual features. Unlike visual signals, speech signals carry most information in their temporal dynamics. This makes auditory scene analysis more difficult. In addition, auditory features are used to identify typical environmental sounds in offices and households, and thus filter out these sounds.

Sound energy, spectral energy vector and Gammatone Frequency Cepstral Coefficients (GFCC) are used mainly because these features are easy to compute, as the system aims to build an efficient scene representation. Furthermore, spectral energy vector and GFCCs are short-term spectral features, which are believed to be the simplest, yet most discriminative features for speaker recognition [Kinnunen and Li, 2009]. In most speaker recognition systems, speaker models need to be trained beforehand by using feature vectors extracted from the same speakers' training utterances [Kinnunen and Li, 2009]. However, no prior knowledge

about the speakers is available for the CAVSA system in this work. Therefore, the task of voice recognition becomes more challenging and a performance deterioration can be anticipated.

In the following, the extraction of these auditory features is introduced. Then these features are tested for discriminating voices and environmental sounds, and the best features are selected.

4.3.1. Spectral energy vector and sound energy

The extraction of spectral energy vector and sound energy is based on the Gammatone Filterbank which has been introduced in section 3.2.1.

In the audio processing (see Fig. 3.2), after a Gammatone Filterbank is applied, the signal envelope for every frequency band is extracted and then a spectral subtraction technique is used to remove the stationary background noise. The resulting signal is denoted as $G(s, f)$ where s and f stand for sample and frequency channel respectively. This Time-Frequency (T-F) representation is called the cochleagram. After an energy-based segmentation, the number of samples L in the segment is computed. Since the sound in one audio segment is produced by the same speaker in our context, $G(s, f)$ is averaged over all samples of the segment:

$$G(f) = \frac{\sum_s G(s, f)}{L}. \tag{4.7}$$

$G(f)$ is called spectral energy vector [Rodemann et al., 2009b] and indicates the distribution of energy over all frequency channels. This feature has been used to differentiate robot directed calls from background noise [Rodemann et al., 2009b].

Sound energy denoted as A is calculated as the sum of $G(f)$ over all frequency channels:

$$A = \sum_f G(f). \tag{4.8}$$

In the same acoustic environment, the louder a sound, the more energy it has. To keep the same form as for the spectral energy vector, A is converted to a population code vector using 100 nodes. The node centers are selected empirically as follows: Each of 10 participants stands in front of the robot and utters 20 natural sentences. It was found that 95% of the audio proto-objects have a sound energy in the range [500, 120000]. Therefore, the node

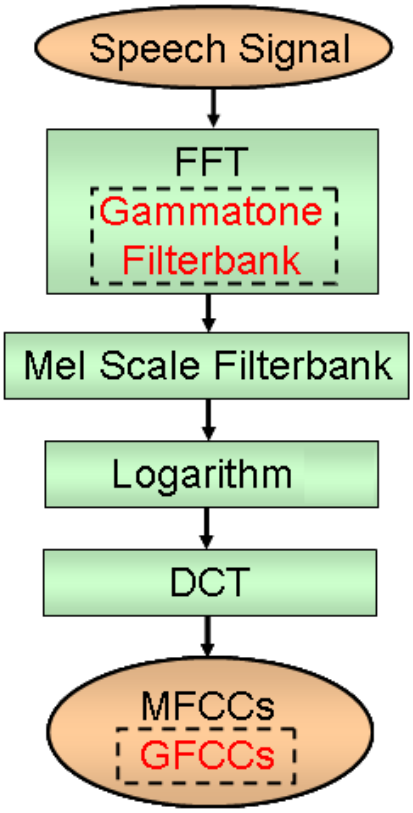


Figure 4.7.: Steps involved in MFCCs (GFCCs) extraction. In case of GFCCs extraction: Gammatone filterbank instead of FFT.

centers are generated as 100 equally spaced points between 500 and 120000. Thus, the population code vector of energy A is defined by a Gaussian function:

$$p_A(n) = \exp\left(\frac{-(n-A)^2}{2 \cdot \delta_n^2}\right), \tag{4.9}$$

where n stands for node center and the standard deviation δ_n is empirically set to 9500.

4.3.2. Gammatone Frequency Cepstral Coefficients

Gammatone Frequency Cepstral Coefficients (GFCC) are extracted based on Mel Frequency Cepstral Coefficients (MFCC), which are perhaps the most popular features and widely used in speaker recognition tasks [Hasan et al., 2004]. As shown in Fig. 4.7, the following steps are commonly performed to derive MFCC features [Ganchev et al., 2005]: First, FFT power spectrum information is derived from a short time windowed segment of speech. Then the Mel-spaced filterbank, a set of triangle filters, are applied to the power spectrum. After

that, the logarithm of all filterbank energies and the DCT of the log filterbank energies are calculated. For standard MFCCs, the DCT coefficients $2-13$ are kept. Usually, the frame energy and delta coefficients are also appended to each feature vector.

The GFCC features described in this work differ from MFCC features only in the power spectrum representation. While MFCC features are based on the FFT power spectrum, the T-F representation in the extraction of GFCC features, the cochleagram $G(s,f)$, is derived from Gammatone filterbank filtering. The cochleagram is analogous to the spectrogram, but provides a higher frequency resolution at low frequencies than at high frequencies [Shao and Wang, 2007]. Since speech signals concentrate more energy at low frequencies and background signals have more energy at higher frequencies [Rodemann et al., 2009b], the performance of GFCC features should be better in speaker recognition than MFCC features. This was verified in [Zhao et al., 2012]. However, it is to mention that the GFCC features, described in [Shao and Wang, 2007] and [Zhao et al., 2012], are not identical to those used in this work. Their GFCC features are not cepstral coefficients, but are derived by applying DCT directly on the feature vector $G(s,f)$ at sample s. Another important reason of using GFCC features is to conveniently extract all auditory features based on the Gammatone filterbank. To obtain a GFCC feature vector for an audio proto-object, the cochleagram $G(s,f)$ is averaged over all samples s in the audio segment, as in Eq. 4.7. In this work, GFCC feature vectors with 13 coefficients (including only the energy term), 16 coefficients, and 26 coefficients (with energy term and their first derivatives) are tested.

4.3.3. Feature Subset Selection

Auditory features (sound energy, spectral energy vector, GFCC features with 13, 16 and 26 coefficients) were introduced in the previous sections. In this section, the performance of these features in discriminating voices and environmental sounds is tested, and the best features are selected. First, auditory features are used for differentiating between 10 sounds produced by 3 women, 5 men, 1 phone bell and 1 toy duck. The subjects spoke natural sentences. 30 audio segments were taken for each sound source, 10 for training of the template and 20 as samples for testing. For each feature, every sample is compared with templates by computing their similarity. The calculation of feature similarity is based on Eq. 4.4. The similarities between samples and templates for each feature form a 10×200- matrix. These matrices are shown in Fig. 4.8, where only the maximum value for each sample is illustrated. In this manner, it is easier to observe if a sample is assigned correctly to its template.

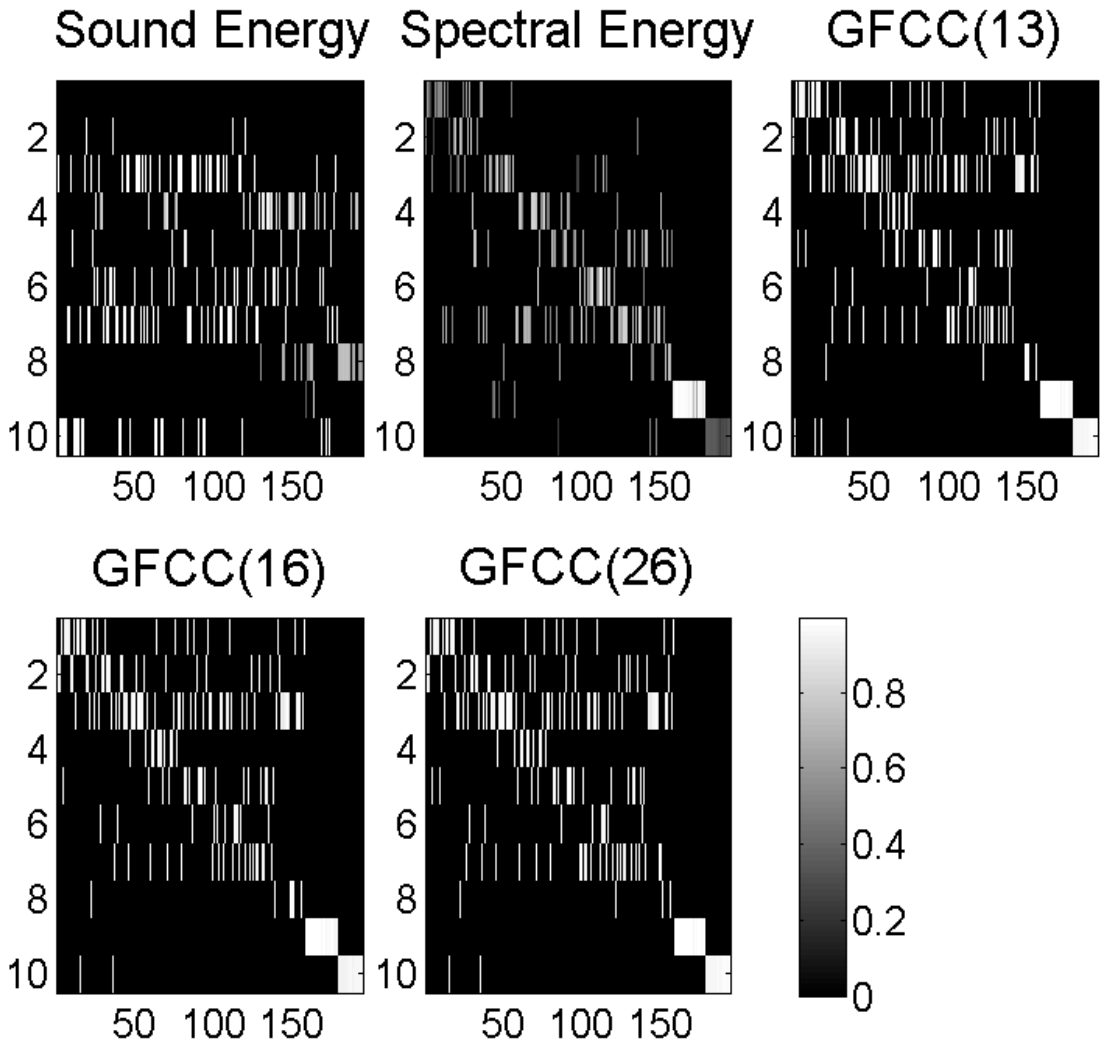


Figure 4.8.: Feature similarity matrices between 10 templates and their samples, where only the maximum value for each sample is shown. The index $1-10$ on the y-axis indicates 10 templates: the template of 3 women, 5 men, a phone bell and a toy duck, respectively. The index $1-200$ on the x-axis indicates 20 samples for each.

Feature	Sound Energy	Spectral Energy	GFCC 13	GFCC 16	GFCC 26
ER	84.5%	43.8%	51%	48%	51%

Table 4.4.: Error Rate (ER) for each auditory feature in the test

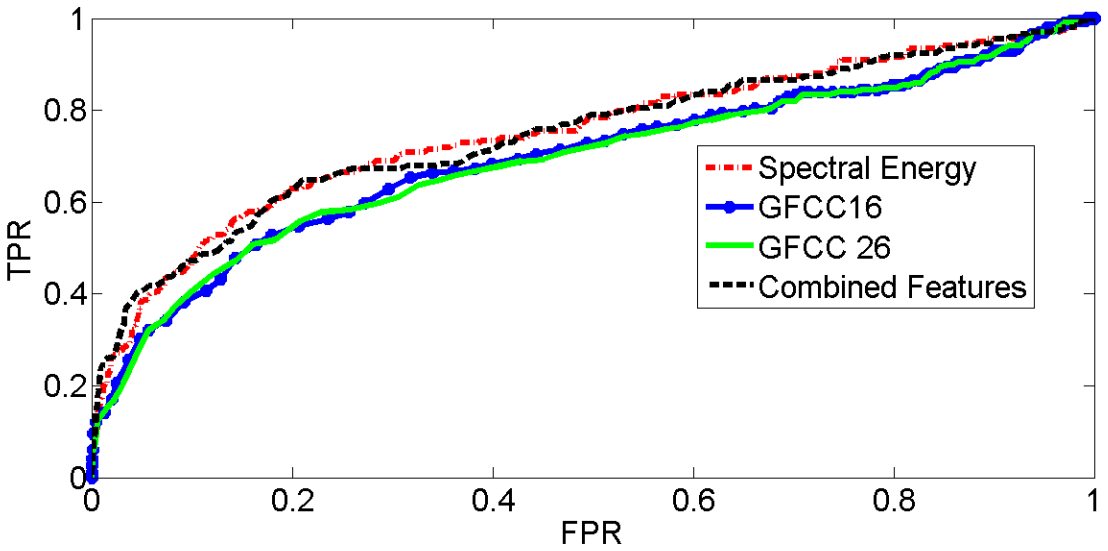


Figure 4.9.: ROC-curves for spectral energy, GFCC with 16 coefficients, GFCC with 26 coefficients and their combination. The point $(0.2104, 0.6482)$ on the curve of combined features has the minimum distance d_m to the perfect classification, $d_m = 0.4099$.

Furthermore, the Error Rate (ER) is computed for each feature:

$$\text{ER} = \frac{\text{total number of falsely assigned samples}}{\text{total number of samples}}.$$

The measured ERs for all feature vectors are shown in Table 4.4. Although ERs of all the auditory features are lower than 90% (ER in case of an assignment by chance), these features are not good enough to reliably distinguish speakers on their own. Therefore, multiple features are combined and the SFFS algorithm (see section 4.2.4) is applied to select a feature subset. The resulting best feature subset contains spectral energy, GFCC with 16 coefficients and GFCC with 26 coefficients. Fig. 4.9 shows the corresponding ROC curves. It can be seen that the performance of the combination is not much better than that of the spectral energy. The minimum distance to the best classification is 0.4099, whereas the minimum distance of the best visual feature subset is only 0.0329 as shown in Fig. 4.6. It is more difficult to identify speakers with auditory features than with visual features especially in the unconstrained settings, because speech signals strongly vary over time. No wonder the performance of most speaker recognition systems is better where complex speaker models are trained beforehand.

As can be seen in Fig. 4.8, although spectral energy and GFCC features perform worse than visual features to differentiate voices, they show a good performance in recognizing the phone bell and toy duck sounds. To verify the performance of auditory features in recognizing environmental sounds, more environmental sounds were tested. Actually, the system does not aim to recognize individual environmental sounds, but does this to filter out these sounds and thus enhance the system robustness. First, sounds of toy ducks, phone bells, door creaking and placing different cups on a table were recorded. Dog barking and cat meowing sounds were downloaded from the internet. 30 audio segments were taken for each kind of sound, 10 for training of the template and 20 as samples for testing. The similarity matrices for auditory features are illustrated in Fig. 4.10. The ERs for each feature vector are measured and shown in Table 4.5. It can be seen that all features, except sound energy,

Feature	Sound Energy	Spectral Energy	GFCC 13	GFCC 16	GFCC 26
ER	69.17%	12.5%	27.5%	24.17%	25%

Table 4.5.: Error Rate (ER) for each auditory feature in the test of environmental sounds

perform much better than in differentiating persons. Therefore, auditory features are applied to identify environmental sounds then filter them.

Features are selected and combined to fulfill the task. While a database for training features to recognize speakers is not available, templates for individual environmental sounds can be trained beforehand. We select different feature subsets for different classes of sounds, because the classification accuracy is often higher than selecting a common feature subset for all classes. Feature subsets are selected using the SFFS algorithm as described in section 4.3.3. The ROCs for selected features are analyzed to set decision thresholds Θ_s for each class. The selected features, Θ_s and the corresponding TPR, FPR for each class of environmental sounds are shown in Table 4.6. A scenario (Scenario IV in section 4.5) will be designed to test the performance of auditory features for recognizing and filtering out environmental sounds.

To summarize, it is hard to recognize voices with auditory features, especially in an unconstrained setting. Therefore, auditory features are not used to group audio proto-objects. They could work in some specific situations though, for example, if there are only two speakers in the scenario and one speaks always louder than the other. In this scenario, even sound energy alone, evaluated as the worst feature (see Table 4.4), is sufficient for differentiating voices. Auditory features were not tested in such specific scenarios though. In most situations, it is risky to use these auditory features to group audio proto-objects in a STM, because once an

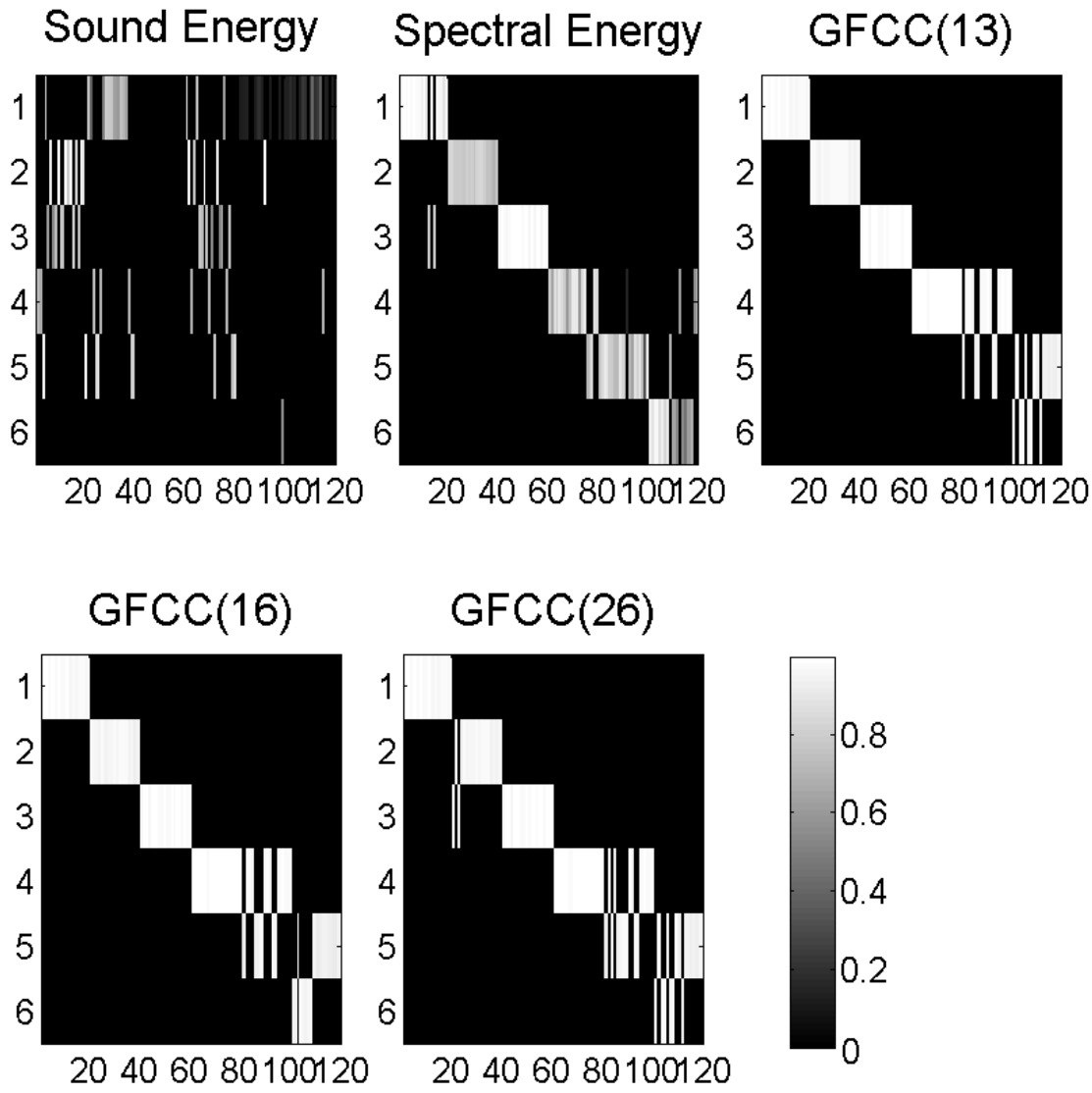


Figure 4.10.: Feature similarity matrices between 6 templates and their samples, where only the maximum value for each sample is shown. The index $1-6$ on the y-axis indicates 6 templates: the feature template of toy ducks, phone bells, door creaking, placing a cup on the table, dog barking and cat meowing. The index $1-120$ on the x-axis indicates their 20 samples for each. There are generally more errors in recognizing the sounds of dog barking and cat meowing, because the sounds of 10 different cats and 12 different dogs are used. These animals differ in ages and emotions. However, it is uncommon that there are so many pets in one household. If sounds of the same animal are used for template and samples, the performance could be better.

Sounds	Θ_s	TPR	FPR	Feature Subset
Toy Duck	0.956	1	0.02	GFCC 26
Phone	0.696	0.95	0.01	spectral energy
Door	0.951	0.95	0.01	spectral energy
Cup	0.98	0.95	0.04	GFCC 26
Dog	0.834	0.85	0.11	spectral energy, GFCC 16
Cat	0.703	0.95	0.07	spectral energy, GFCC 26

Table 4.6.: Decision threshold Θ_s, TPR,FPR and selected features for classes of each individual environmental sound: sounds of toy ducks, phone bells, door creaking, placing different cups on the table, dog barking and cat meowing.

audio proto-object is falsely assigned in the STM, the system is not able to correct the error by itself. Hence an auditory STM is not applied in this work. On the other hand, auditory features show a good performance in identifying typical environmental sounds. In order to filter out these sounds, the spectral energy vector and GFCC features with 16 and with 26 coefficients are selected and stored in audio proto-objects.

4.4. Setting of CAVSA for Dialogue Scenarios

CAVSA is used to assign words to corresponding speakers in natural dialogue scenarios as shown in Fig. 4.11. CAVSA is set up in the following.

4.4.1. Proto-Objects

According to the analysis of visual features in the previous section, the system collects Hair LBP, Clothes LBP, Face LBP, height as well as the position evidence vector $p_v(\theta)$ (as in equation 4.3) in one visual proto-object. An audio proto-object contains start time, segment length, the mean energy and the position evidence vector $p_a(\theta)$. The extraction of these auditory features was introduced in section 3.2.1. In order to filter out environmental sounds, the spectral energy vector, and GFCC features with 16 and 26 coefficients are additionally stored in the audio proto-object, as explained in section 4.3.3.

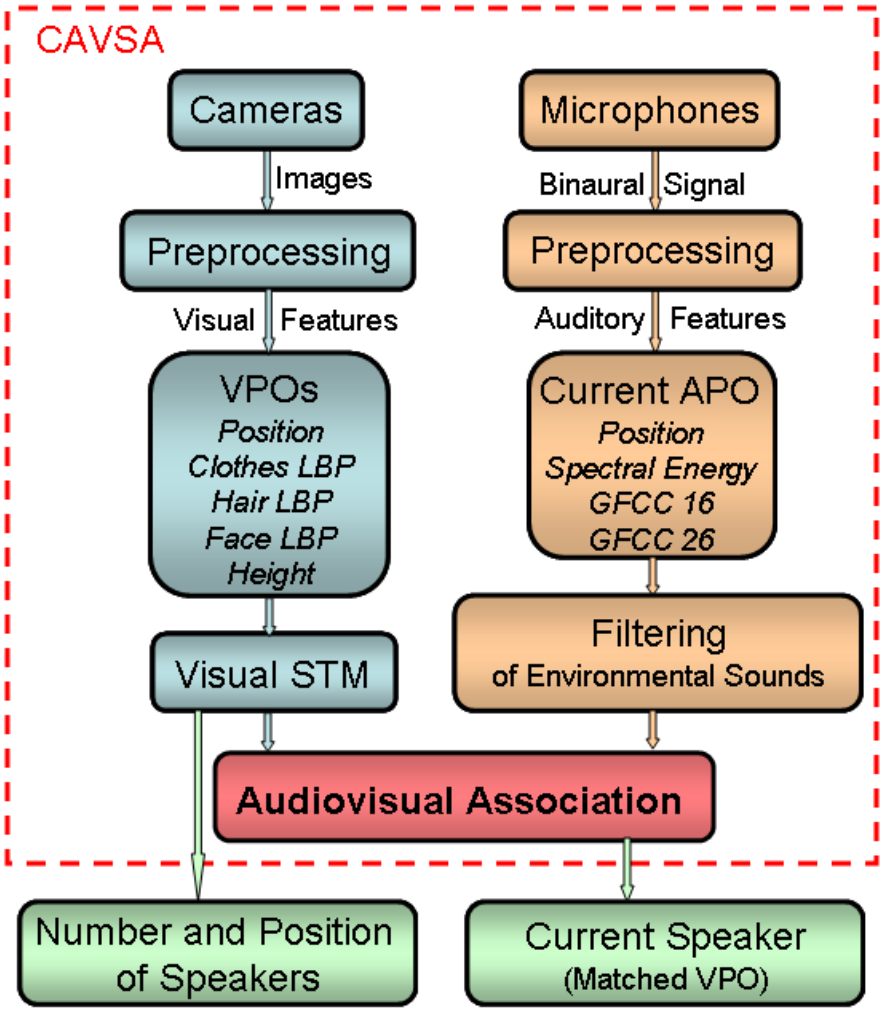


Figure 4.11.: System architecture of CAVSA for dialogue scenarios. CAVSA aims at learning the number and positions of interacting partners, as well as who is currently speaking. Additional visual features (Clothes LBP, Hair LBP, Face LBP and Height) are applied to recognize persons while additional auditory features (spectral energy vector, GFCC features with 16 and 26 coefficients) are used to filter out environmental sounds. APO: audio proto-object, VPO: visual proto-object.

4.4.2. Visual STM

The visual STM is used to group visual proto-objects from the same speaker and memorize proto-objects, even if the corresponding persons are out of sight for a while due to the person's or robot's movement or the failure of the face detection. The procedure of inserting an incoming proto-object into the STM is described in section 2.3. In this section, the "compare" step is configured.

Let $X_m = (x_1, x_2, ...x_F)$ $(m \in [1, M])$ denote the m-th incoming proto-object and $Y_n = (y_1, y_2, ...y_F)$ $(n \in [1, N])$ denote the n-th proto-object in STM , where $x_i \in X_m$ (resp. $y_i \in Y_n$) stands for the i-th $(i \in [1, F])$ feature in X_m (resp. Y_n), F is the number of features in a proto-object, M the number of current incoming proto-objects and N the number of proto-objects in STM. First, the similarity s_i between feature vectors x_i and y_i is computed based on equation 4.4. Next, the similarity S_{mn} between proto-objects X_m and Y_n is measured using the mean operator:

$$S_{mn} = \frac{\sum_{i=1}^{F} s_i \cdot w_i}{\sum_{i=1}^{F} w_i}, \tag{4.10}$$

where w_i, the weight of the i-th feature, is a time decay function as defined in equation 2.2. Since the position feature has a stronger variation than other features when people move, a smaller time constant $\tau = 5\ s$ is set for the position feature. For other features, τ is set equal to 100 s.

4.4.3. Audiovisual Association

The current audio proto-object A is linked to a visual proto-object V_j $(j \in [1, N])$ from the STM based on the similarity of their position evidence vectors p_a and p_{v_j}, to find the current speaker. First, the auditory position evidence vector p_a is extended using linear interpolation to have the same length as the visual position vector p_{v_j}. The interpolated auditory vector is denoted as p'_a. Then the probability that A and V_j are caused by the same person is computed as the similarity of p_{v_j} and p'_a, as follows:

$$P_{common}(A, V_j) = \frac{p'_a \cdot p_{v_j}}{||p'_a||\,||p_{v_j}||}. \tag{4.11}$$

The CAVSA system also considers the situation where the current sound is related to a per-

son who has not yet been seen, as in Algorithm 3 ("Audiovisual association: Consideration of unseen VPOs"). Chapter 3 already proved that Algorithm 3 has better performance than Algorithm 2 with the basic approach. When the related visual source has not yet been spotted, Algorithm 3 was able to detect the situation and refused audiovisual integration, while the basic approach in Algorithm 2 selected a wrong visual proto-object for audiovisual association.

Let V_{N+1} denote the proto-object of the unseen person. The probability that the current audio proto-object A is associated with V_{N+1} is described by:

$$P_{common}(A, V_{N+1}) = \frac{\sum_\theta U(\theta) \cdot p'_a(\theta)}{\sum_\theta p'_a(\theta)}, \tag{4.12}$$

where $U(\theta)$ is a view memory defined in equation 3.3. The system then searches for the visual proto-object V_l which has maximum probability $P_{common}(A, V_j)$ ($j \in [1, N]$) as in equation 2.3. If $P_{common}(A, V_{N+1})$ is larger than $P_{common}(A, V_l)$, the matched visual proto-object is considered to have not yet been seen. Otherwise, the audio proto-object is associated with visual proto-object V_{jMax}.

In this application, the system must answer the question "who is currently speaking ?" using audiovisual association independent of whether the association is reliable or not. Therefore, the uncertainty of the audiovisual association based on equation 3.6 is not calculated. However, the uncertainty information could be used according to requirements and applications. For example the uncertainty information could be sent to the dialogue manager in a dialogue system. If an audiovisual association is unreliable, a dialogue manager may ask for confirmation.

4.5. Results

To evaluate the performance of CAVSA, experiments were carried out in a small room (for details see chapter 3). Three dialogue scenarios were designed as shown in Fig. 4.12 to test if the CAVSA system using additional visual features is able to correctly determine the number and position of speakers, as well as the current speaker. Person 1, 2 and 3 are represented with red, blue and black icons. Red, blue and black dashed lines stand for motion trajectories of person 1, 2 and 3, respectively. Each scenario was repeated five times with different participants to verify results. To ease the description, the repetitions of

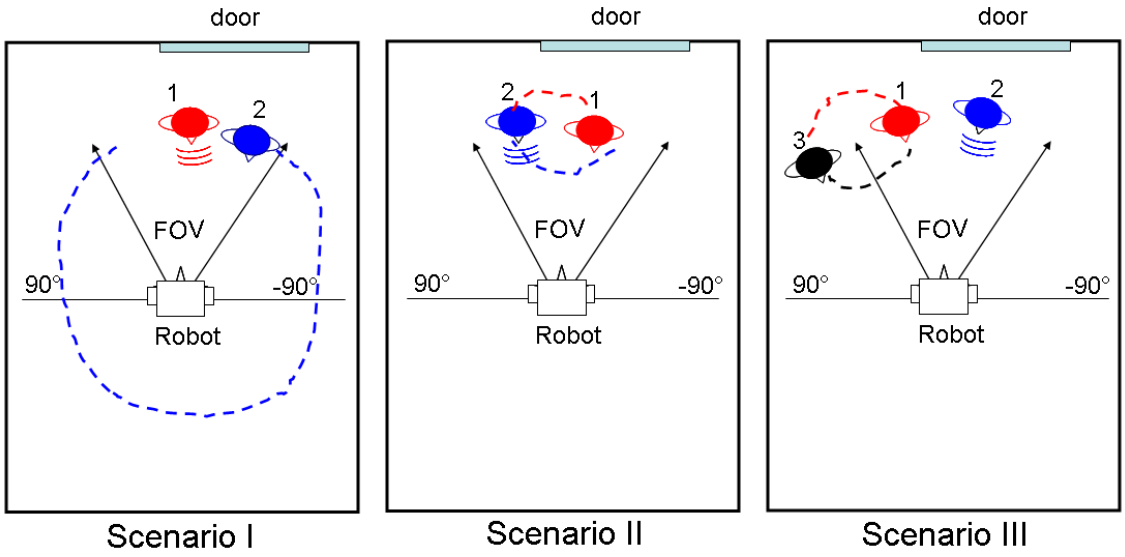


Figure 4.12.: Sketch of three scenarios. **Scenario I**: two participants talk to the robot at the beginning. Then speaker 2 leaves the FOV, passes behind the robot and returns; **Scenario II**: two speakers change their positions; **Scenario III**: person 1 and person 2 talk to the robot, while person 3 is outside the FOV and remains silent. Then person 1 and person 3 exchange their positions. In all scenarios, the people in the FOV talk alternatingly.

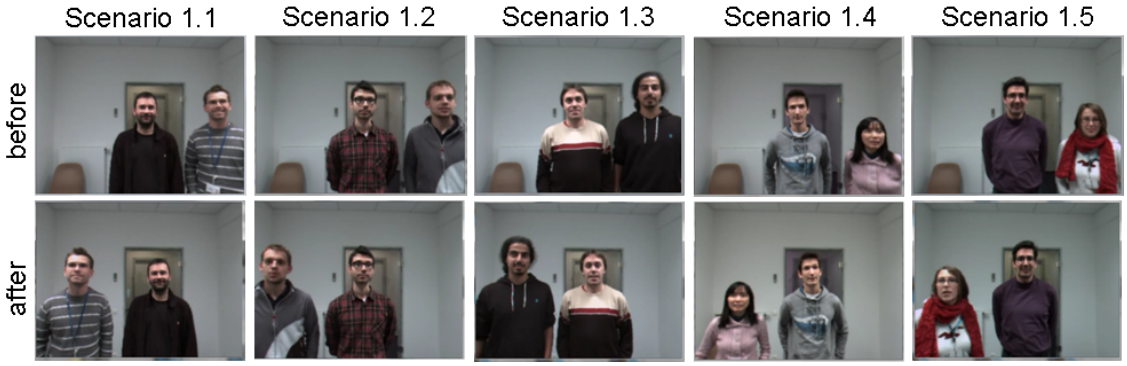


Figure 4.13.: Scenario I.1- I.5: participants in the camera FOV before and after the movement.

Scenario I are called Scenario I.1, Scenario I.2, ..., Scenario I.5, the repetitions of Scenario II Scenario II.1, Scenario II.2, ..., Scenario II.5, and the repetitions of Scenario III Scenario III.1, Scenario III.2, ..., Scenario III.5. Five groups of participants for three scenarios are shown in Fig. 4.13, Fig. 4.14 and Fig. 4.15 before and after the corresponding scenario motion. Threshold Θ_s for the similarity S_{mn} (in equation 4.10) is set to 0.7 based on data from Scenarios I.1, II.1 and III.1.

For each scenario, the CAVSA system using position and visual features (Hair LBP, Height, Clothes LBP and Face LBP) is compared with the previous version described in chapter 3 which used only position to group visual proto-objects in STM. In the following the system using only position is denoted as "*CAVSA-Pos*" and the version using multiple visual features as "*CAVSA-Multi*". *CAVSA-Pos* behaves similarly to methods in [Bennewitz et al., 2005,

Figure 4.14.: Scenario II.1- II.5: participants in the camera field of view before and after the movement.

Figure 4.15.: Scenario III.1- III.5: participants in the camera FOV before and after the movement.

Fritsch et al., 2003, Yan et al., 2013b], given people unknown to the system beforehand.

Furthermore, an additional scenario (Scenario IV) was designed to test the performance of auditory features for filtering out environmental sounds. In this scenario, a person talked to the robot, then several background sounds were produced such as sounds of placing the cup on the table, phone bells and door creaking.

The results of these four scenarios are described below.

4.5.1. Results for Scenario I

In Scenario I.1, the starting configuration had person 1 at around 0° azimuth angle and person 2 at -25°. They talked to the robot alternatingly. At time step 205, person 2 moved out of the FOV and passed behind the robot. During this time only person 1 spoke. At time step 356, person 2 reappeared at around 25° and both talked to the robot alternatingly again. Fig. 4.16 (1) and (2) show the result of scene representation over time using *CAVSA-Pos* and *CAVSA-*

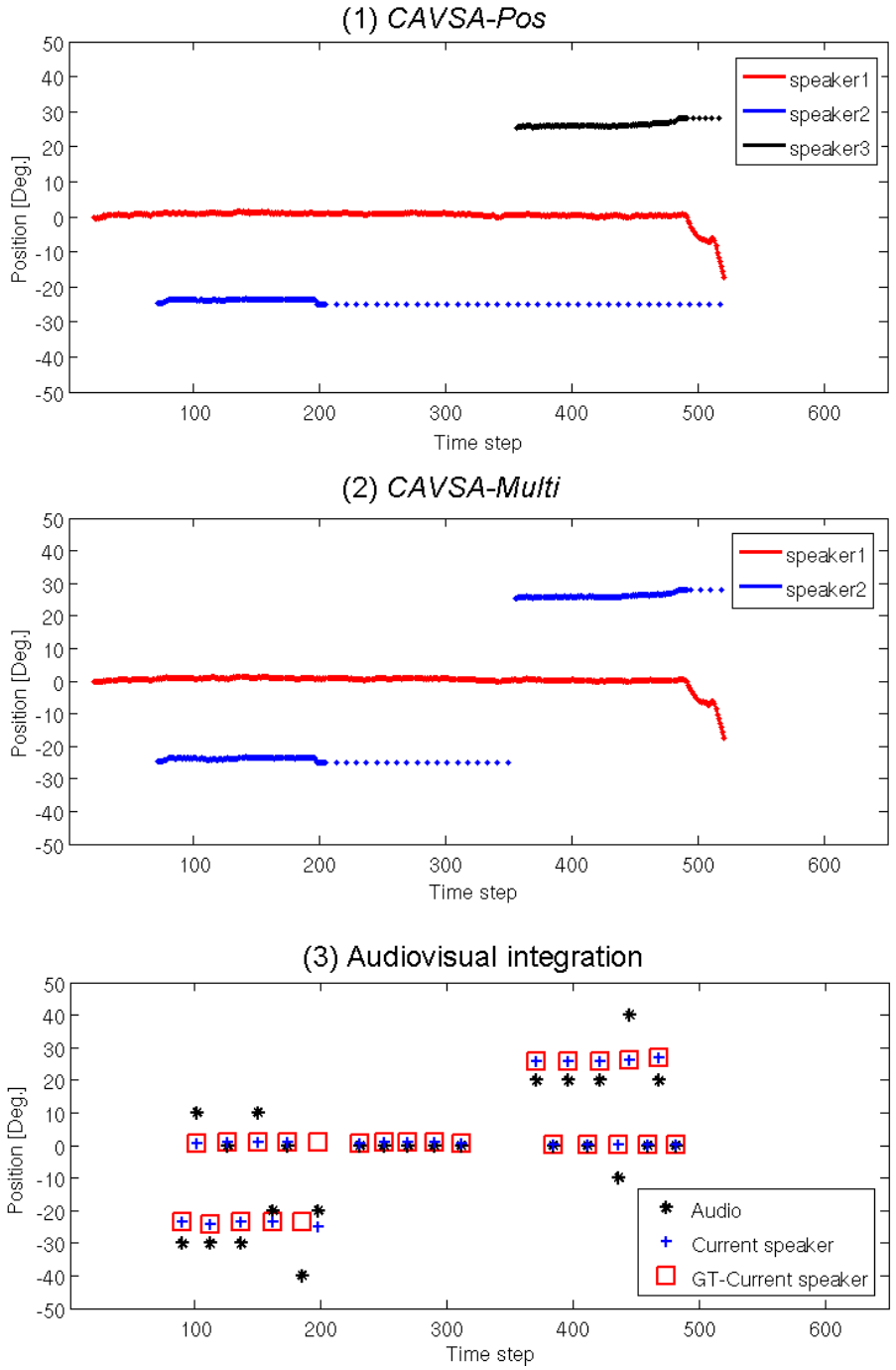


Figure 4.16.: Scenario I: (1) visual scene representation with *CAVSA-Pos*; (2) visual scene representation with *CAVSA-Multi*; (3) result of the audiovisual association. In (1) and (2), person identity is visualized using different colors. Red, blue and black are used for person 1, person 2 and person 3, respectively. In (3), "Audio" represents sound position, "Current Speaker" the visual position of the detected current speaker and "GT-Current Speaker" the visual position of the true current speaker. Ground truth data was added manually.

Multi, respectively. Person identity is visualized using different colors. Red, blue and black are used for person 1, person 2 and person 3, respectively. A solid line means that the related person is currently in the FOV, while dashed lines indicate disappeared persons. The cameras record images at 10 frame/s, that is one time step is 0.1 s. Fig. 4.17 (for Scenario II.1) and Fig. 4.18 (for Scenario III.1) visualize results in the same manner. As can be seen in Fig. 4.16 (1) and (2), when person 2 left the FOV, both *CAVSA-Pos* and *CAVSA-Multi* stored the corresponding visual proto-objects in the STM. However, *CAVSA-Pos* considered person 2 as a new person ("person 3"), when he disappeared briefly and returned from a different azimuth angle, because similarity between the new and saved position was too small. In comparison, *CAVSA-Multi* was able to correctly represent the visual scene, since it integrated also visual features. Similarly, *CAVSA-Multi* represented the visual scene correctly with $\Theta_s = 0.7$ in Scenario I.2 -I.5, while *CAVSA-Pos* considered person 2 as a new person.

Fig. 4.16 (3) visualizes the result of audiovisual association based on position information. *CAVSA-Multi* was used here for grouping visual proto-objects in STM. "Audio" stands for sound position, while "Current Speaker" represents the visual position of the detected current speaker and "GT-Current Speaker" the visual position of the true current speaker. Ground truth data was annotated by hand. We can see that audiovisual association worked most of the time. But the current speaker was considered to have not been seen at time step 185 and was falsely detected at time step 198. Both errors were due to inaccurate sound localization.

4.5.2. Results for Scenario II

In Scenario II.1, person 1 stood initially at around -10° azimuth angle and person 2 at 20°. They talked to the robot alternatingly. At time step 280, speaker 1 moved from -10° to 20° and speaker 2 from 20° to -20°. Then they continued to talk in the same manner. Fig. 4.17 shows the result of the scene representation using *CAVSA-Pos* and *CAVSA-Multi*, as well as the result of the audiovisual association. *CAVSA-Pos* interpretation of the situation was that speaker 2 did not move, speaker 1 exited and a new speaker ("person 3") came into the scenario, while *CAVSA-Multi* correctly represented the scene over time. Fig. 4.17 (3) shows that only one error of audiovisual association occurred at time step 157 also due to an erroneous sound localization. The changes in the scene can be correctly detected in the remaining repetitions of Scenario II except for Scenario II.4, where the male participant is recognized as a new person after changing his position. Since he did not stand completely in the FOV, the similarity of his Clothes LBP feature before and after changing position is very low and similarity S_{mn} is lower than $\Theta_s = 0.7$.

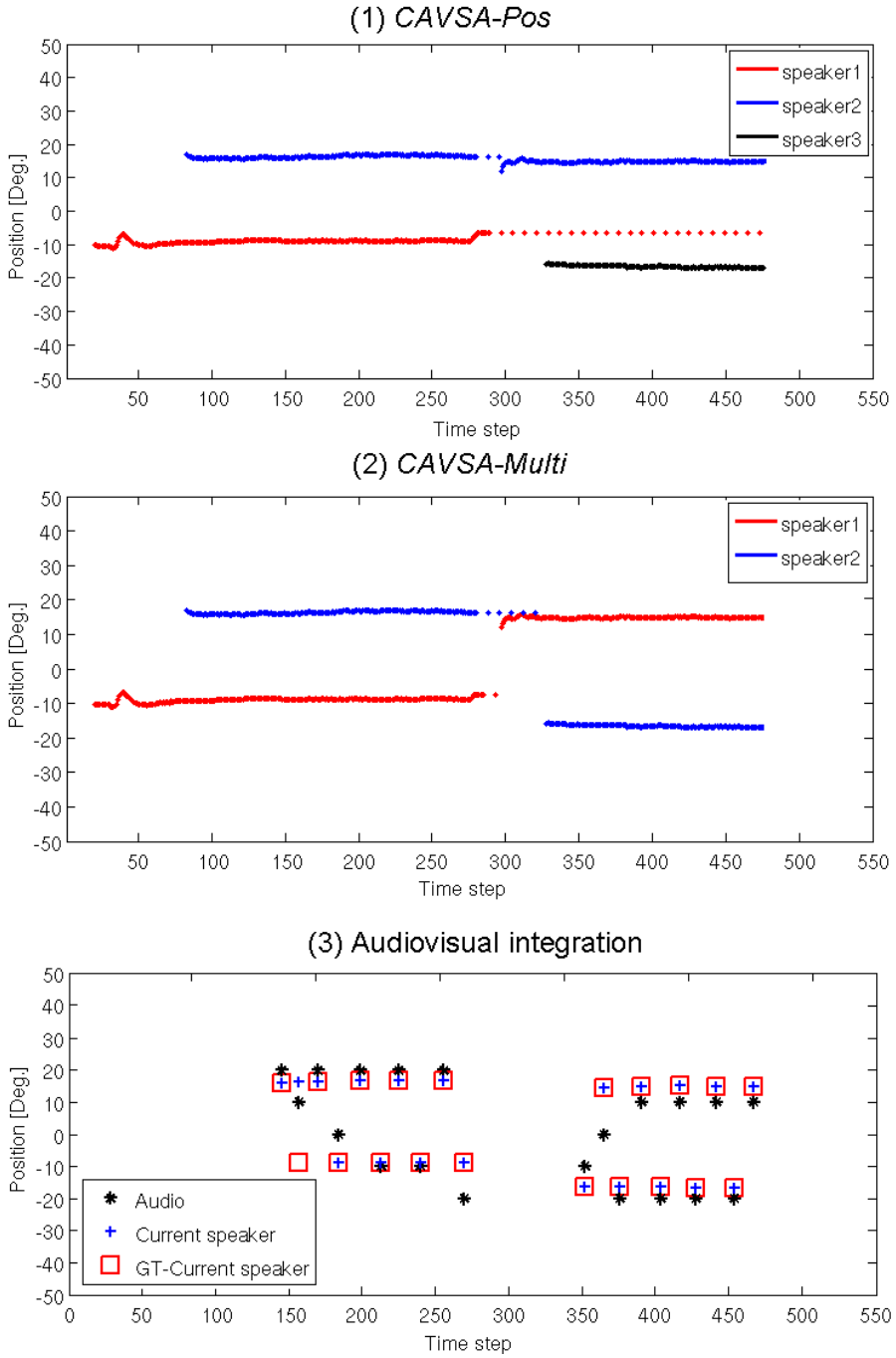


Figure 4.17.: Scenario II: (1) visual scene representation with *CAVSA-Pos*; (2) visual scene representation with *CAVSA-Multi*; (3) result of the audiovisual association. In (1) and (2), person identity is visualized using different colors. Red, blue and black are used for person 1, person 2 and person 3, respectively. In (3), "Audio" represents sound position, "Current Speaker" the visual position of the detected current speaker and "GT-Current Speaker" the visual position of the true current speaker. Ground truth data was added manually.

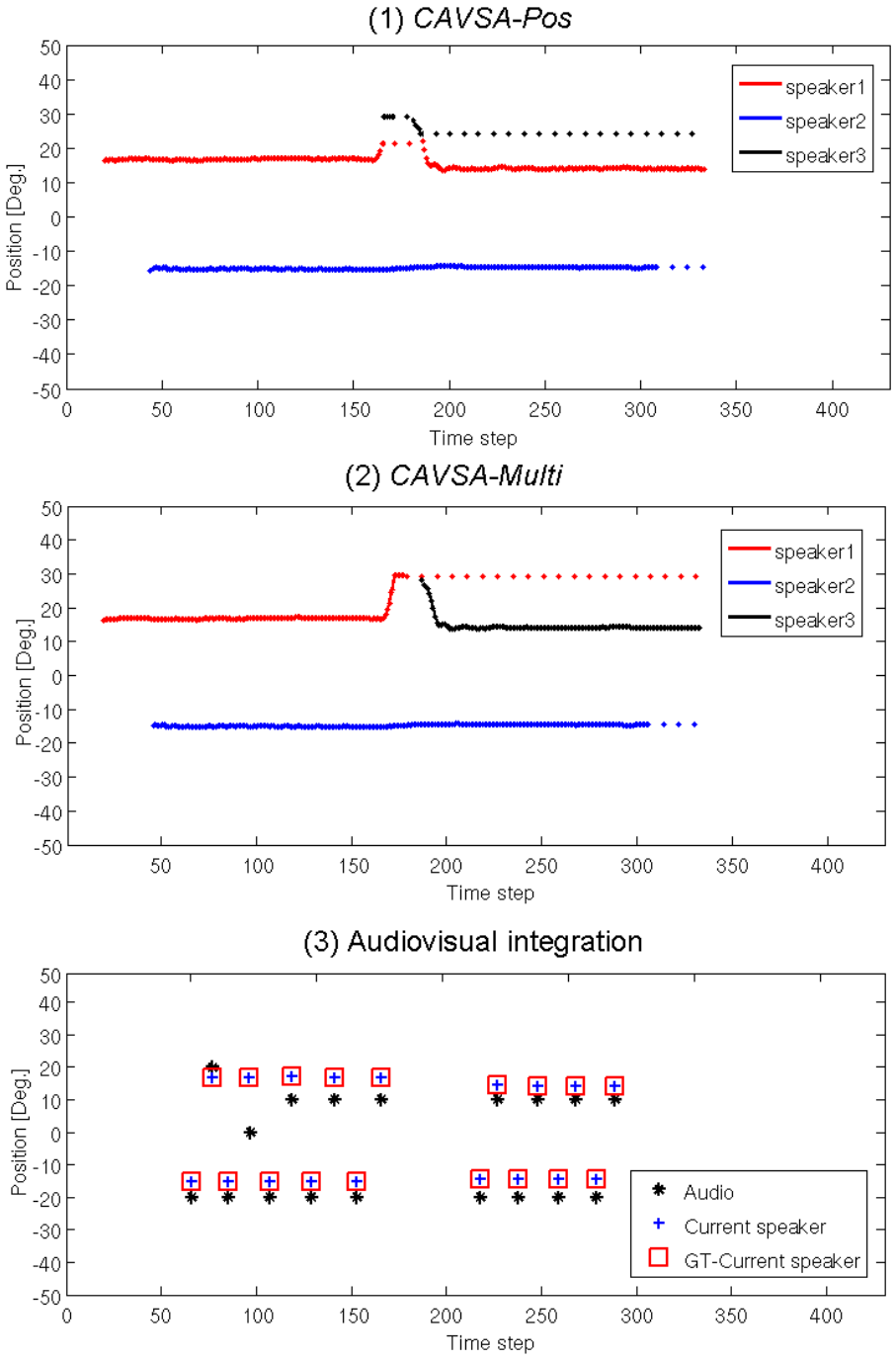


Figure 4.18.: Scenario III: (1) visual scene representation with *CAVSA-Pos*; (2) visual scene representation with *CAVSA-Multi*; (3) result of the audiovisual association. In (1) and (2), person identity is visualized using different colors. Red, blue and black are used for person 1, person 2 and person 3, respectively. In (3), "Audio" represents sound position, "Current Speaker" the visual position of the detected current speaker and "GT-Current Speaker" the visual position of the true current speaker. Ground truth data was added manually.

4.5.3. Results for Scenario III

In Scenario III.1, person 1 stood initially at around 17° azimuth angle, person 2 at −15° and person 3 was outside of the FOV. Person 1 and person 2 talked to the robot alternatingly. At time step 173, person 1 exited at 30° (the border of the FOV). Person 3 appeared at 30° at time step 187 and moved to 15° at time step 195. Then person 2 and person 3 talked to the robot alternatingly. As can be seen from Fig. 4.18 (1), *CAVSA-Pos* detected speaker 1 as a new person "person 3" at time step 171 because of his fast movement. Then person 1 exited from the border of the FOV and speaker 3 appeared from the same side, so that they were considered as the same person ("person 3"). At time step 190, person 3 moved to the old position of person 1, and was detected as person 1, while "person 3" was detected to be out of the FOV. As shown in Fig. 4.18 (2), *CAVSA-Multi* represented the dynamic scene correctly. In Scenario III.2 and III.3, person 1 and 3 were recognized as the same person by *CAVSA-Multi* with $\Theta_s = 0.7$, since they have similar visual features (Hair LBP, Height, Clothes LBP and Face LBP). It was found that once a visual proto-object is falsely assigned, the system is not able to correct the error by itself. Fig. 4.18 (3) illustrates that all the audiovisual associations were successful.

Since the same type of errors of audiovisual association was found in multiple scenarios, the error rate (falsely assigned APOs/ all APOs) is calculated based on all 15 scenarios. The total error rate is 7%, i.e. words are most of the time correctly assigned to the matched speakers.

4.5.4. Results for Scenario IV

In Scenario IV, a person talked to the robot, then placed a cup on the table. After that, his mobile phone rang and he opened the door to go outside to answer the call and closed the door behind him. In this process, 14 audio proto-objects were generated by the system, 6 for utterances of the person, 1 for placing the cup on the table, 5 for phone bells and 2 for door creaking. The templates of the environmental sounds are trained beforehand as explained in section 4.3. Auditory features that are used to train each template and decision thresholds are shown in Table 4.6. The results of filtering environmental sounds are illustrated in Fig. 4.19. The ground truth data of audio proto-objects was annotated by hand. We can see that different environmental sounds are correctly recognized and thus filtered out.

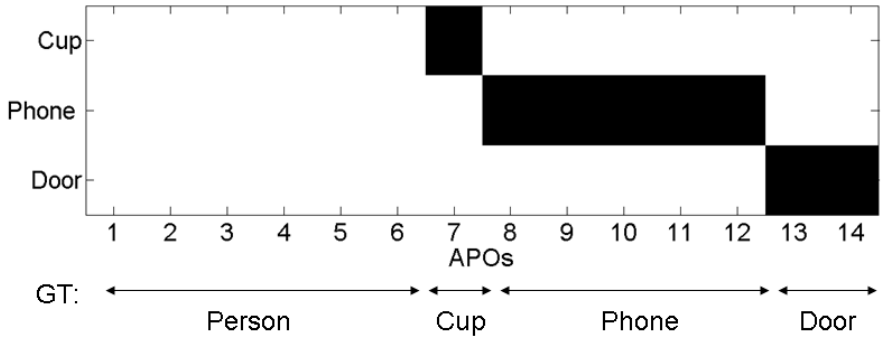


Figure 4.19.: Scenario IV: result of filtering environmental sounds that are the sounds of placing the cup on the table, phone bells and door creaking. GT: Ground truth data that was added manually.

4.6. Discussion

In this chapter, CAVSA is applied in human-robot dialogue scenarios. It has been shown that the CAVSA system was able to correctly determine the number and position of speakers as well as who is currently speaking in most real-world scenarios. The previous version in chapter 3 and state-of-the-art methods can not recognize a person again when he leaves the scene for a while. This is because they use only position information to group visual proto-objects in the STM. In contrast, the system integrating more visual features, was shown to be able to recognize reappearing persons most of the time in more complicated scenarios where speakers left and entered the scene, or switched positions. Audiovisual association sometimes failed if sound location was falsely estimated. But the current speaker was correctly detected in a large majority of cases.

Furthermore, auditory features are applied to improve the robustness of the system. It has been shown that they had good performance to recognize typical environmental sounds and thus filter out these sounds.

5. Summary

This work has presented a system of computational audiovisual scene analysis (CAVSA). Assuming that all meaningful sounds originate from human speakers, the CAVSA system is able to learn how many speakers are in the scenario, where the speakers are and who is currently speaking in most dialogue scenarios. Speakers are unknown in advance.

For CAVSA, auditory and visual features are first collected in form of proto-objects, which are understood here as a compressed form of various features. Audio and visual proto-object are used together in one framework at the first time. In this manner, dense visual images and sparse sound sources are transformed into a common representation, which eases the audiovisual integration. Proto-objects are designed for operation in realistic environments, since they can be tracked, pointed or referred to without identification and enable a flexible interface to behavior-control in robotics. In CAVSA, proto-objects for the same speaker are grouped together in auditory and visual STMs using position information. STMs help to memorize persons, even if they disappear from the field of view for a while. Finally, visual and audio proto-objects of the same speaker are integrated. The system links the current sound to the corresponding speaker in multi-person environments, and measures the uncertainty of the audiovisual integration. Additionally, I consider the situation where the current sound is related to a person who has not yet been seen.

CAVSA has many potential applications. In this thesis, it is applied in two applications. First, CAVSA is tested in online adaptation of audio-motor maps. It enables the system to find the matched visual source to the current audio source in multi-person environments, thus vision can be used to provide precise position information in online adaptation of audio-motor maps. In comparison to state-of-the-art methods, the proposed system does not need control over the robot motion and thus can work during other tasks. In terms of learning progress, it was shown that the approach with CAVSA was more robust in multi-person scenarios. It was also shown that the performance of online adaptation does not degrade when the current speaker has not yet been seen. Furthermore, the system was able to bootstrap with a randomized audio-motor map in multi-person environments. In case of hardware modifications,

the adaptation process was capable of quickly restoring the old localization performance in the single-person scenario. In multi-person scenarios, the recovery period depended on how large the modification was made.

In online adaptation of audio-motor maps, only a small set of simple features is sufficient, since it is not necessary to identify persons. However, the original version of CAVSA sometimes fails in dialogue scenarios, for example when a person greatly alters his position outside the field of view. This is because only position information is used to group visual proto-objects in the STM. Therefore I employed more visual features to recognize persons for improving the capability of CAVSA and tested the performance in real-world dialogue scenarios. It was shown that the CAVSA system was able to assign words to corresponding speakers in a large majority of cases. A speaker was recognized again when he left and entered the scene, or changed his position even with a newly appearing person. Furthermore, a set of simple auditory features was tested. They had good performance to recognize typical environmental sounds in offices and households. Therefore, auditory features were used to filter out these sounds.

Outlook

Possibilities for improving the performance of CAVSA are proposed in the following.

The current system can recognize persons most of the time using a set of simple visual features. The fixed set of features is selected and a fixed similarity threshold is set, based on a given database. However, the recognition sometimes fails, because other visual features might be better than the selected features in some scenarios. To solve this problem, an adaptive feature selection method that depends on specific scenarios should be used in the future.

Furthermore, although the current speaker is detected correctly in most scenarios, audiovisual association sometimes fails if sound localization is falsely estimated. Moreover, when a person changes his position outside the camera field of view and talks to the robot, his visual proto-object stored in the STM, especially the position feature, can not be updated. In this case, his audio proto-object can not be linked to his visual proto-object using position information. A possible solution for above problems is using more auditory features to recognize voices. Once an audio and a visual proto-object are associated, the association can be main-

tained for a certain time. Imagine a simple dialogue scenario with a robot and 2 speakers wearing different colored clothes. The person in red clothes always speaks louder than the other in blue clothes. The system links "red" to "louder". Then, when the louder voice is heard, the robot recognizes that the person in the red clothes is talking, even if the person is invisible. However, it was shown in section 4.3.3 that the performance of the current auditory features is not good enough for differentiating voices. Therefore, it could be better to find a set of more powerful auditory features. Alternatively, an adaptive feature selection method could select features among the current auditory features dependent on scenarios. For example, only sound energy is sufficient for differentiating voices in the above given scenario.

There is a number of issues I do not tackle in this work and which need to be addressed in future work. First, I 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, the current system employed filters for audio proto-objects based on simple cues that remove most environmental sounds. Second, the system should be able to detect if a speaker talks to the robot or to other persons. Head orientation and/or the view direction of the eyes could be detected to overcome this issue by assuming that the people talking to the robot looks at its face [Kaminski et al., 2006, Wallhoff et al., 2006]. Third, since gaze control plays an important role in scene representation, robot gaze could be controlled to direct fixation towards ROIs which are defined depending on tasks. This involves in the topic of active vision [Backer et al., 2001, Yann et al., 2008]. For example, the current robot system can detect the situation that the current sound is related to a persons who has not yet been seen. Then the robot could turn its head to the possibly unknown person to update the scene representation. In certain tasks, the robot could be asked to track a certain person among multiple people to keep him always in the camera FOV.

A. Abbreviations

APO Audio Proto-Object

ASA Auditory Scene Analysis

BSS Blind Source Separation

CASA Computational Auditory Scene Analysis

CAVSA Computational AudioVisual Scene Analysis

ER Error Rate

FFT Fast Fourier Transform

FOV Field of View

FPR False Positive Rate

GFB Gammatone FilterBank

GFCC Gammatone Frequency Cepstral Coefficients

HRI Human Robot Interaction

HRTF Head Related Transfer Function

IID Interaural Intensity Difference

ITD Interaural Time Difference

LBP Local Binary Pattern

NIP Normalized Inner Product

PDF Probability Density Function

PO Proto-Object

ROC Receiver Operating Characteristic

ROI Region of Interest

SC Superior Colliculus

SFFS Sequential Forward Floating Selection

SFS Sequential Forward Selection

STM Short-Term Memory

T-F Time-Frequency

TPR True Positive Rate

VAD Voice Activity Detection

VPO Visual Proto-Object

B. Audio-Motor Map

Humans determine where a sound is coming 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 [Rodemann et al., 2008]. 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) [Blauert, 2001] as shown in Fig. B.1. 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 [Rodemann et al., 2007]. 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 [Rodemann et al., 2007]. 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 [Davis et al., 2003, Gill et al., 2000, Wanrooij and Opstal, 2004, Rodemann, 2011], which are 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) [Viste and Evangelista, 2004]. For a set of audio cues it provides the motor command such as head motion in

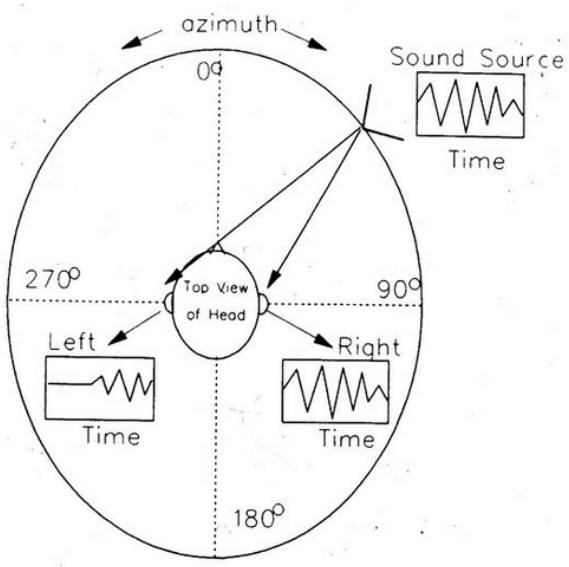


Figure B.1.: IID and ITD are two most widely used cues for localizing a sound source. The sound arriving at the ear that is furthest from the sound source is delayed (time difference) and lower in amplitude (intensity difference). Image originated from: http://www.cns.nyu.edu

the form of an azimuth and an elevation angle to focus on the estimated position of the sound source. We concentrate on audio-motor maps for azimuth angle 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 as [Rodemann et al., 2008, Rodemann, 2010], 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) [Blauert, 2001]:

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

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. [Viste and Evangelista, 2004] added a scaling factor α_f to Eq. B.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:

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

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

C. Sound Localization with Population-Coded Cues

In this thesis, population-coded audio cues as proposed in [Rodemann et al., 2009a] are used in sound localization. In a population coding, feature values are indirectly represented by the activity of a number of nodes (neurons) [Pouget et al., 2000]. The measurement of audio cues is explained in [Heckmann et al., 2006]. Since audio cues, IID and ITD, depend on the frequency spectrum, they are measured for each frequency channel f. Frequency channel index f is in the range from $(1-100)$, corresponding to Gammatone center frequencies between 100 Hz and 11000 Hz. 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. 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:

$$C^i(f,n) = \exp\left(\frac{-(c^i(f)-n)^2}{2\cdot\delta^2}\right), \tag{C.1}$$

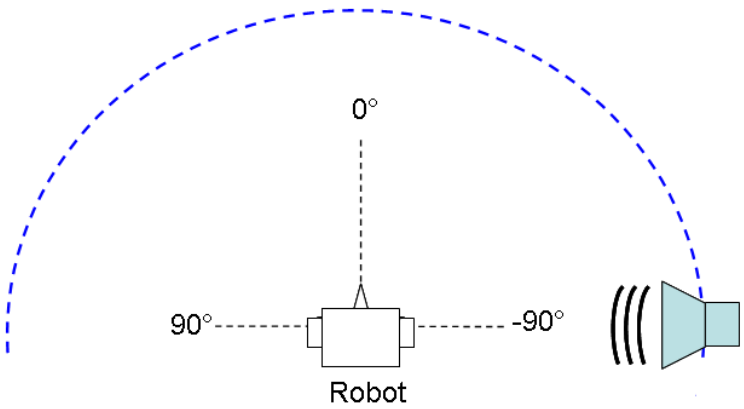


Figure C.1.: An example of sound source at $-90°$ in azimuth angle.

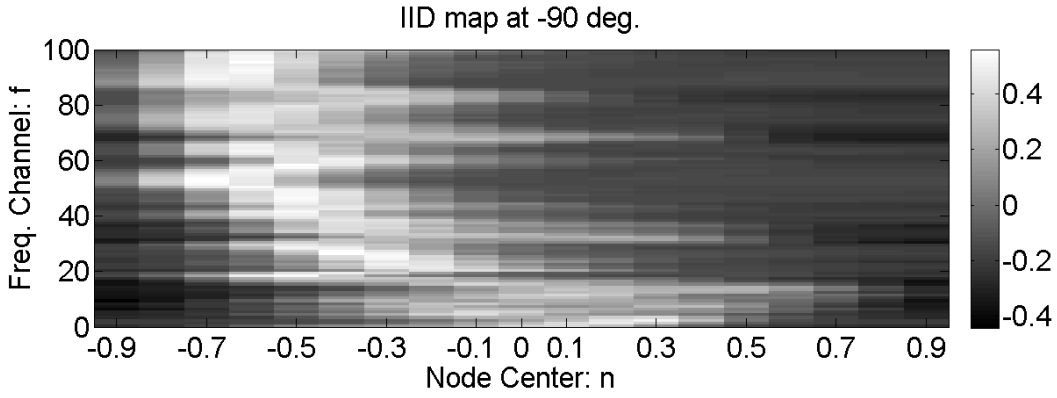


Figure C.2.: An example of the IID map at $-90°$, corresponding to the population code vectors of measured IID we would expect if a sound source were at position $\theta = -90°$. The vector in each frequency channel is normalized to mean 0 and norm 1.

where n stands for node center and the standard deviation δ is set to 0.1. For example when a sound source is at $-90°$ as illustrated in Fig. C.1, the expected population code vectors of measured IID $C^1(f,n)$ is shown in Fig. C.2. All measurements are added over time for an audio proto-object and the sum is denoted as $C'^i(f,n)$. In this manner, more than one stimulus can be encoded in a population coding and uncertainties can be nicely represented. For example if there are different values of IID due to noise or echoes, the probabilities for all candidate values can be stored. The combined cue measurements over the complete segment of an audio proto-object enable a better localization performance [Rodemann et al., 2009a].

In order to map audio cues to azimuth position candidates, an audio-motor map is used. The audio-motor map in this thesis is represented as a simple look-up table $M^i_\theta(f,n)$ ($i = 1$ 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 θ. Only azimuth positions θ between $-90°$ and $90°$ are taken into account because of the mechanical constraint of the robot head.

Fig. C.2 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 $M^i_\theta(f,n)$ and measured cues $C'^i(f,n)$, we compare the population response $C'^i(f,n)$ with $M^i_\theta(f,n)$ for all positions θ by computing scalar products to acquire the position evidence vector W^i_θ. The peak in the position vector W_θ ($W_\theta = \sum_{i=1}^{2} W^i_\theta$) is taken as the estimated sound source position. Fig. C.3 shows an example of a position evidence vector with the peak at $40°$.

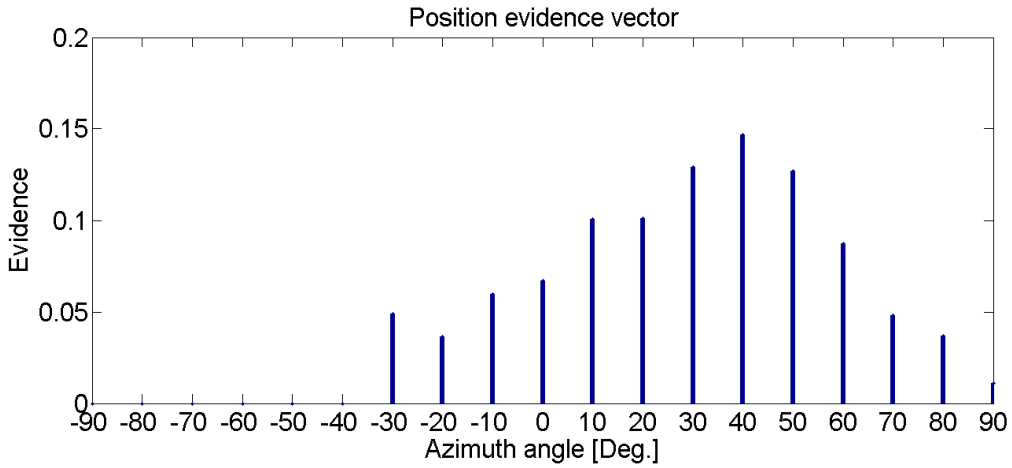


Figure C.3.: An example of a position evidence vector. The vector is normalized to sum 1. Here the estimated azimuth is 40°.

List of Publications by the Author

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

Yan, R., Rodemann, T., and Wrede, B. (2011b). Learning of audiovisual integration. In *Proceedings of International Conference on Development and Learning,* volume 2, pages $1-7$.

Yan, R., Rodemann, T., and Wrede, B. (2012a). Simple auditory and visual features for human-robot dialog scene analysis. In *Proceedings of the International Conference on Intelligent Robots and Systems,* pages $700-706$.

Yan, R., Rodemann, T., and Wrede, B. (2012b). Audiovisual Integration in Dialog Scenarios. In *Sixth International Conference on Cognitive and Neural Systems*.

Yan, R., Rodemann, T., and Wrede, B. (2013a). Computational audiovisual scene analysis in online adaptation of audio-motor maps. *IEEE Transactions on Autonomous Mental Development,* $5(4):273-287$.

Bibliography

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262.

Atkinson, R. C. and Shiffrin, R. M. (1971). The control processes of short-term memory. Technical report, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Backer, G., Mertsching, B., and Bollmann, M. (2001). Data- and model-driven gaze control for an active-vision system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1415–1429.

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., and Behnke, S. (2005). Enabling a humanoid robot to interact with multiple persons. In *Proceedings of the International Conference on Dextrous Autonomous Robots and Humanoids*.

Blauert, J. (2001). *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT press, 3rd edition.

Bolder, B., Brandl, H., Heracles, M., Janssen, H., Mikhailova, I., Schmüdderich, J., and Goerick, C. (2008). Expectation-driven autonomous learning and interaction system. In *Proceedings of the International Conference on Humanoid Robots*.

Bolder, B., Dunn, M., Gienger, M., Janssen, H., Sugiura, H., and Goerick, C. (2007). Visually guided whole body interaction. In *Proceedings of the International Conference on Robotics and Automation*, pages 3054 –3061.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press.

Brown, G. J. and Wang, D. (2005). *Speech enhancement*, chapter Separation of Speech by Computational Auditory Scene Analysis, pages 371–402. Springer.

Bui, T. H. (2006). Multimodal dialogue management - State of the art. Technical report, Human Media Interaction Department, University of Twente.

Bulkin, D. A. and Groh, J. M. (2006). Seeing sounds: Visual and auditory interactions in the brain. *Current Opinion in Neurobiology*, 16(4):415–419.

Burr, D. and Alais, D. (2006). *Progress in Brain Research*, volume 155, chapter Combining visual and auditory information, pages 243–258. Elsevier.

Casanovas, A. L., Monaci, G., and Vandergheynst, P. (2007). Blind audiovisual source separation using sparse representations. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 301–304.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307.

Checka, N., Wilson, K., Rangarajan, V., and Darrell, T. (2003). A probabilistic framework for multi-modal multi-person tracking. In *Conference on Computer Vision and Pattern Recognition Workshop*, page 100.

Cichocki, A. and Amari, S. (2002). *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons.

Corcoran, P., Costache, G., and Mulryan, R. (2005). Automatic indexing of consumer image collections using person recognition techniques. In *Proceedings of the International Conference on Consumer Electronics*, pages 127 – 128.

Davis, K. A., Ramachandran, R., and May, B. J. (2003). Auditory processing of spectral cues for sound localization in the inferior colliculus. *Journal of the Association for Research in Otolaryngology*, 4(2):148–163.

DeBello, W. M. and Knudsen, E. I. (2004). Multiple sites of adaptive plasticity in the owl's auditory localization pathway. *The Journal of Neuroscience*, 24(31):6853–6861.

Divenyi, P. (2004). *Speech separation by humans and machines*. Springer.

Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., and Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 102, pages 16907–16908.

Finger, H., Ruvolo, P., Liu, S., and Movellan, J. R. (2010). Approaches and databases for online calibration of binaural sound localization for robotic heads. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 4340–4345.

Fisher, J. W. and Darrell, T. (2004). Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413.

Friedland, G., Yeo, C., and Hung, H. (2009). Visual speaker localization aided by acoustic models. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 195–202.

Fritsch, J., Kleinehagenbrock, M., Lang, S., Fink, G. A., and Sagerer, G. (2004). Audiovisual person tracking with a mobile robot. In *Proceedings of the International Conference on Intelligent Autonomous Systems*, pages 898–906.

Fritsch, J., Kleinehagenbrock, M., Lang, S., Plötz, T., Fink, G. A., and Sagerer, G. (2003). Multi-modal anchoring for human-robot interaction. *Robotics and Autonomous Systems*, 43(2-3):133–147.

Ganchev, T., Fakotakis, N., and Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. In *Proceedings of the 10th International Conference on Speech and Computer*, volume 1, pages 191–194.

Gates, B. (2007). A robot in every home. *Scientific American*, 296:58–65.

Gill, D., Troyansky, L., and Nelken, I. (2000). Auditory localization using direction-dependent spectral information. *Neurocomputing*, 32-33:767–773.

Gold, J. I. and Knudsen, E. I. (2000). Abnormal auditory experience induces frequency-specific adjustments in unit tuning for binaural localization cues in optic tectum of juvenile owls. *The Journal of Neuroscience*, 20(2):862–877.

Haider, F. and Moubayed, S. A. (2012). Towards speaker detection using lips movements for human-machine multiparty dialogue. In *Proceedings of FONETIK*, pages 117–120.

Hasan, M., Jamil, M., and Rahman, M. (2004). Speaker identification using mel frequency cepstral coefficients. In *Proceedings of the International Conference on Electrical and Computer Engineering*, pages 565–568.

Heckmann, M., Berthommier, F., and Kroschel, K. (2002). Noise adaptive stream weighting in audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, 2002(11):1260–1273.

Heckmann, M., Rodemann, T., Joublin, F., and Goerick, C. (2006). Auditory inspired binaural robust sound source localization in echoic and noisy environments. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 368 – 373.

Hershey, J. and Movellan, J. (1999). Audio-vision: Using audio-visual synchrony to locate sounds. In *Proceedings of the Conference Neural Information Processing Systems*, volume 12, pages 813–819.

Hollingworth, A. and Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28(1):113–136.

Hörnstein, J., Lopes, M., Santos-Victor, J., and Lacerda, F. (2006). Sound localization for humanoid robots-building audio-motor maps based on the HRTF. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 1170–1176.

Hung, H. and Friedl, G. (2008). Towards audio-visual on-line diarization of participants in group meetings. In *Proceedings of the European Conference on Computer Vision*, pages 1–12.

Hyde, P. S. and Knudsen, E. I. (2002). The optic tectum controls visually guided adaptive plasticity in the owl's auditory space map. *Nature*, 415:73–76.

IFR (2012). World robotics 2012 service robots. Technical report, International Federation of Robotics, http://www.ifr.org/service-robots/statistics/.

Iwano, K., Yoshinaga, T., Tamura, S., and Furui, S. (2007). Audio-visual speech recognition using lip information extracted from side-face images. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(1):064506.

Jaffre, G., Joly, P., Université, I., and Sabatier, P. (2004). Costume: A new feature for automatic video content indexing. In *Proceedings of the 7th RIAO Conference: Coupling approaches, coupling media and coupling languages for information retrieval*, pages 314–325.

Jain, A. and Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153 – 158.

Kacelnik, O., Nodal, F., Parsons, C., and King, A. (2006). Training-induced plasticity of auditory localization in adult mammals. *Public Library of Science Biology*, 4(4):627–638.

Kaminski, J. Y., Shavit, A., Knaan, D., and Teicher, M. (2006). Head orientation and gaze detection from a single image. In *Proceedings of the International Conference Of Computer Vision Theory And Applications*, pages 85–92.

Karaoguz, C., Rodemann, T., Wrede, B., and Goerick, C. (2013). Learning information acquisition for multi-tasking scenarios in dynamic environments. *IEEE Transactions on Autonomous Mental Development*, 5(1):46–61.

Khalidov, V., Forbes, F., Hansard, M., Arnaud, E., and Horaud, R. (2008). Audio-visual clustering for multiple speaker localization. In *Proceedings of the 5th International Workshop on Machine Learning for Multimodal Interaction*, volume 5237, pages 86 – 97.

Kim, H., Komatani, K., Ogata, T., and Okuno, G. (2007). Auditory and visual integration based localization and tracking of humans in daily-life environments. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 2021 – 2027.

King, A. J., Hutchings, M. E., Moore, D. R., and Blakemore, C. (1988). Developmental plasticity in the visual and auditory representation in the mammalian superior colliculus. *Nature*, 332(6159):73–76.

Kinnunen, T. and Li, H. (2009). An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 52(1):12–40.

Knudsen, E. I. (1998). Early blindness results in a degraded auditory map of space in the optic tectum of the barn owl. *In Proceedings of the National Academy of Sciences of the USA*, 85(16):6211–6214.

Knudsen, E. I. (2002). Instructed learning in the auditory localization pathway of the barn owl. *Nature*, 417(6886):322–328.

Knudsen, E. I. and Knudsen, P. F. (1989). Vision calibrated sound localization in developing barn owls. *The Journal of Neuroscience*, 9(9):3306–3313.

Knudsen, E. I. and Knudsen, P. K. (1985). Vision guides the adjustment of auditory localization in young barn owls. *Science*, 230(4725):545–548.

Marcel, S., Rodriguez, Y., and Marcel, G. H. (2006). On the recent use of local binary patterns for face authentication. Technical report, Idiap Research Institute.

Markov, K. (2009). Advanced approaches to speaker diarization of audio documents. In *Proceedings of Joint Conferences on Pervasive Computing*, pages 179–184.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The psychological review*, 63(2):81–97.

Nakadai, K., Hidai, K., Mizoguchi, H., Okuno, H. G., and Kitano, H. (2001). Real-time auditory and visual multiple-object tracking for humanoids. In *Proceedings of 17th International Joint Conference on Artificial Intelligence*, pages 1425–1436.

Nakashima, H. and Mukai, T. (2005). 3D sound source localization system based on learning of binaural hearing. In *Proceedings of the International Conference on Systems, Man and Cybernetics*, volume 4, pages 3534–3539.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Heinze, H. J., and Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *Neuroscience*, 27(42):11431–11441.

Noulas, A. and Krose, B. J. A. (2007). On-line multi-modal speaker diarization. In *Proceedings of the International Conference on Multimodal interfaces*, pages 350–357.

Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59.

Okuno, H. G., Nakadai, K., Lourens, T., and Kitano, H. (2004). Sound and visual tracking for humanoid robot. *Applied Intelligence*, 20(3):253–266.

Park, D., Park, J., and J.H.Han (1999). Image indexing using color histogram in the cieluv color space. In *Proceedings of the 5th Japan-Korea Joint Workshop on Computer Vision*, pages 126–132.

Paul M. Hofman, J. G. V. R. . A. J. V. O. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, 1(5):417–421.

Pedersen, M. S., Larsen, J., Kjems, U., and Parra, L. C. (2008). *Springer Handbook on Speech Processing*, chapter A Survey of Convolutive Blind Source Separation Methods, pages 1065–1084. Springer.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nature Neuroscience Reviews*, 1(2):125–132.

Pudil, P., Novovicová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1-2):127–158.

Ramírez, J., Górriz, J. M., and Segura, J. C. (2007). *Robust Speech Recognition and Understanding*, chapter Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, pages 1–22. I-TECH Education and Publishing.

Rodemann, T. (2010). A study on distance estimation in binaural sound localization. In *Proceedings of the conference on Intelligent Robots and Systems*, pages 425–430.

Rodemann, T. (2011). Spectral cues to source position in robots with arbitrary ear shapes. In *Proceedings of the International Conference on Advanced Robotics*, pages 453–458.

Rodemann, T., Heckmann, M., Schölling, B., Joublin, F., and Goerick, C. (2006a). Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 860 – 865.

Rodemann, T., Ince, G., Joublin, F., and Goerick, C. (2008). Using binaural and spectral cues for azimuth and elevation localization. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 2185 – 2190.

Rodemann, T., Joublin, F., and Goerick, C. (2006b). Continuous and robust saccade adaptation in a real-world environment. *KI-Künstliche Intelligenz*, 06(3):23–26.

Rodemann, T., Joublin, F., and Goerick, C. (2009a). Audio proto objects for improved sound localization. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 187 – 192.

Rodemann, T., Joublin, F., and Goerick, C. (2009b). Filtering environmental sounds using basic audio cues in robot audition. In *Proceedings of the International Conference on Advanced Robotics*, pages 1 – 6.

Rodemann, T., Karova, K., Joublin, F., and Goerick, C. (2007). Purely auditory online-adaptation of auditory-motor maps. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 2015 – 2020.

Rosenthal, D. F. and Okuno, H. G., editors (1998). *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, Inc.

Sanchez-Riera, J., Alameda-Pineda, X., Wienke, J., Deleforge, A., Arias, S., Čech, J., Wrede, S., and Horaud, R. (2012). Online multimodal speaker detection for humanoid robots. In *Proceedings of IEEE International Conference on Humanoid Robots*, pages 126 – 133.

Schettini, R., Ciocca, G., and Zuffi, S. (2001). *Color Imaging Science: Exploiting digital media*, chapter A Survey Of Methods For Colour Image Indexing And Retrieval In Image Databases. Media, John Wiley.

Schmüdderich, J. M. (2010). *Multimodal Learning of Grounded Concepts in Embodied Systems*. Phd thesis, Faculty of Technology, Bielefeld University.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656.

Shao, Y. and Wang, D. (2007). Incorporating auditory feature uncertainties in robust speaker identification. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 277–280.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, 12(5):182–186.

Slaney, M. (1993). An efficient implementation of the patterson-holdsworth auditory filter-bank. Technical report, Apple Computer Company.

Stein, T., Peelen, M. V., and Sterzer, P. (2011). Adults' awareness of faces follows newborns' looking preferences. *PLoS ONE*, 6(12):e29361.

Stricker, M. A. and Orengo, M. (1995). Similarity of color images. In *Proceedings of SPIE Storage and Retrieval for Image and Video Databases III*, volume 2420, pages 381–392.

Swain, M. J. and Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

Topi, M., Timo, O., Matti, P., and Maricor, S. (2000). Robust texture classification by subsets of local binary patterns. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 3, pages 935 – 938.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518.

Visscher, K. M., Kaplan, E., Kahana, M. J., and Sekuler, R. (2007). Auditory short-term memory behaves like visual short-term memory. *PLoS Biology*, 5(3):e56.

Viste, H. and Evangelista, G. (2004). Binaural sound localization. In *Proceedings of the International Conference on Digital Audio Effects*, pages 145–150.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., and Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, 158(2):252–258.

Wallace, M. T. and Stein, B. E. (2006). Early experience determines how the senses will interact. *Journal of Neurophysiology*, 97(1):921–926.

Wallhoff, F., Ablaßmeier, M., and Rigoll, G. (2006). Multimodal face detection, head orientation and eye gaze tracking. In *Proceedings of the International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 13–18.

Wang, D. and Brown, G. J., editors (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press/Wiley-Interscience.

Wanrooij, M. M. V. and Opstal, A. J. V. (2004). Contribution of head shadow and pinna cues to chronic monaural sound localization. *The Journal of Neuroscience*, 24(17):4163–4171.

Weisswange, T. H., Rothkopf, C. A., Rodemann, T., and Triesch, J. (2011). Bayesian cue integration as a developmental outcome of reward mediated learning. *PLoS ONE*, 6(7):e21575.

Willert, V., Eggert, J., Adamy, J., Stahl, R., and Körner, E. (2006). A probabilistic model for binaural sound localization. *IEEE Transactions on Systems, Man and Cybernetics*, 36(5):982–994.

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

Yan, R., Rodemann, T., and Wrede, B. (2011b). Learning of audiovisual integration. In *Proceedings of International Conference on Development and Learning*, volume 2, pages 1–7.

Yan, R., Rodemann, T., and Wrede, B. (2012). Simple auditory and visual features for human-robot dialog scene analysis. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 700–706.

Yan, R., Rodemann, T., and Wrede, B. (2013a). Computational audiovisual scene analysis in online adaptation of audio-motor maps. *IEEE Transactions on Autonomous Mental Development*, 5(4):273–287.

Yan, R., Tee, K. P., Chua, Y., Huang, Z., and Li, H. (2013b). An attention-directed robot for social telepresence. In *Proceedings of the 1st International Conference on Human-Agent Interaction, III-1-2*.

Yann, D., Shahram, B., Luc, D., and François, C. (2008). Gaze control of an active vision system in dynamic scenes. In *Workshop on Vision in Action: Efficient strategies for cognitive agents in complex environments*.

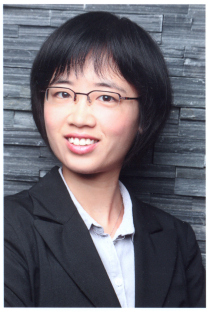
Yost, W. A., Popper, A. N., and Fay, R. R., editors (2008). *Auditory Perception of Sound Sources*. Springer.

Zhao, X., Shao, Y., and Wang, D. (2012). CASA-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1608 – 1616.

Zweig, M. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577.

Zwiers, M. P., Opstal, A. J. V., and Cruysberg, J. R. M. (2001). A spatial hearing deficit in early-blind humans. *Neuroscience*, 21(9):1–5.

Curriculum Vitae



Rujiao Yan was born on April 16, 1981, in Henan Province, China.

She received the Diploma degree in Electronical Engineering from the Technical University Dortmund, Germany, in 2008. From 2010 to 2013, she has been working toward the Ph.D. degree at the Research Institute for Cognition and Robotics (CoR-Lab) in Bielefeld University, Germany, and was a guest scientist at the Honda Research Institute Europe (HRI-EU) in Offenbach/Main, Germany. Her research interests were human-robot interaction, multimodal perception and biologically inspired robots. During this time, the work presented in this thesis has been carried out.