Computational Audiovisual Scene Analysis for Dialog Scenarios

Rujiao Yan, Tobias Rodemann, Britta Wrede

2011

Preprint:

This is an accepted article published in IROS 2011 Workshop on Cognitive Neuroscience Robotics. The final authenticated version is available online at: https://doi.org/[DOI not available]

Computational Audiovisual Scene Analysis for Dialog Scenarios

Rujiao Yan[1,2], Tobias Rodemann[2] and Britta Wrede[1]

Abstract— **We introduce a system for Computational Audio-Visual Scene Analysis (CAVSA) with a focus on human-robot dialogs in multi-person environments. The general target of CAVSA is to learn who is speaking now, where the speaker is, and whether the speaker is talking to the robot or to other persons. In the application specified in this paper, we aim at estimating the number and position of speakers using several auditory and visual cues. Our test application for CAVSA is the online adaptation of audio-motor maps, where vision is used to provide position information about the speaker. The system can perform this adaptation during the normal operation of the robot, like when the robot is engaging in conversation with a group of humans. Comparing our online adaptation of audio-motor maps using CAVSA with prior online adaptation methods, our approach is more robust in situations with more than one speaker and when speakers dynamically enter and leave the scene.**

I. Introduction

In most robotics scenarios, a robot is usually interacting with multiple people. Thus it should be able to learn who is speaking now, where the speaker is, and whether the speaker is talking to the robot or to other persons. Computational Audiovisual Scene Analysis (CAVSA) is aimed at fulfilling these tasks. CAVSA plays an important role in human-robot interaction, for instance it enables the robot to better understand dialog situations, improves speech recognition by assigning words to speakers, and relates visual and auditory features of a speaker. To evaluate the performance of CAVSA we employ it for the online adaptation of audio-motor maps. This task depends strongly on a correct scene representation and the performance can be measured by comparison to audio-motor maps calibrated in standard offline procedures.

In robotics, many sound localization systems use audio-motor maps, which describe the relationship between binaural cues and sound position in motor coordinates (azimuth and elevation). The main binaural cues are interaural time difference (ITD) and interaural intensity difference (IID) [1]. Using audio-motor maps one can compute the sound source position from measured audio cues. We concentrate on audio-motor maps for azimuth to ease the description of our algorithm, but the approach can be expanded to elevation. Audio-motor maps can be calibrated offline by measuring audio cues for several known positions [2]. However, audio-motor maps can change and need to be relearned whenever any relevant part of the robot or the environment was modified, for example, microphone type, microphone position, robot head and room. Additionally, it is difficult to estimate the quality of the current maps. Hence a continuous online adaptation has been considered during the normal operation of the robot [3]. It is known that humans continuously adapt the audio-motor maps to their current auditory periphery, while the dimensions of the head and external ears are growing from birth to adulthood. Even adults have sound localization plasticity, for instance when molds are placed into the external ears to alter audio-motor maps [4]. Rodemann et al. [3] suggested a purely auditory online adaptation approach, where audio provides position information of limited precision.

Vision plays an important role in calibration of audio-motor maps in humans and animals [5]–[7]. Thus vision has been used as the feedback signal for higher precision in adaptation of audio-motor maps in robotics, as per [8], [9]. However, the approach in [8] fails when more than one person or the wrong speaker appears in the camera image. Nakashima et al. [9] proposed another approach using visual feedback in a simplified environment, where a red marker was attached to the sound source and no other red object exists. These methods employ heuristics for linking visual and auditory information and can only work in limited environments. Besides, both methods need extra head motions to search for the visual marker. In comparison to state of the art, our CAVSA method is used to find the correct visual correspondence of the current sound source, and aims at enabling online adaptation to run in more complex environments. If CAVSA selects an unrelated visual signal for the adaptation, the quality of audio-motor maps may deteriorate. Given precise measurements of visual position and audio cues, the quality of maps depends on the performance of CAVSA. This is the reason why we test CAVSA in online adaptation of audio-motor maps. Our system does not require specific motions of the robot, so that audio-motor maps can be continuously online adapted during the normal operation of the robot.

For CAVSA the scene is represented with auditory and visual features using the concept of proto-objects. Proto-objects can combine an arbitrary number of features in a compressed form. For more information about proto-objects see [10], [11]. The visual and audio proto-objects for the same speaker are then integrated based on position information.

A. Comparison to related work

Currently there is a broad range of applications using audiovisual integration in multi-person environments. The first application is speaker recognition (see e.g. [12]), where

1 Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, 33594 Bielefeld, Germany {ryan,bwrede}@cor-lab.uni-bielefeld.de

2 Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073 Offenbach, Germany tobias.rodemann@honda-ri.de

an audiovisual database consisting of all speakers is required for training. The second application is audiovisual multi-person tracking. Most methods use only sound position as an auditory feature and thus fail when a speaker leaves the scenario for a while and reappears or the speaker moves while not talking [13], [14]. In this case CAVSA could flexibly add more auditory and visual features to identify the speaker. Multi-person tracking can also be implemented in a smart-room environment [15], where many auditory and visual sensors are installed. The third group is audiovisual speaker diarization systems [16], which can index who spoke when in a video file. The training of diarization is normally an offline process, where the data can only be processed after complete recording. Another application searches for the visual part of the current speaker in a video using synchrony between lip motion and speech [17]. The approach requires that the lip of the current speaker is always in the field of view. In comparison to these methods, our CAVSA approach can combine many low or mid level auditory and visual features to achieve a high performance of audiovisual integration. It runs in an unsupervised, real-time, online and incremental manner. Besides, we use only a humanoid robot head with a pair of cameras and a pair of microphones. In this work just the left camera is employed to capture the visual signal. The head is mounted on a pan-tilt unit.

Additionally, while current approaches in Computational Auditory Scene Analysis (CASA) mostly deal with parallel sources using microphone arrays (see e.g. [18]), we focus on purely sequential sounds. Our system could be used after sound source separation in case of concurrently active sources.

II. CAVSA

In this section we introduce the concept of Computational AudioVisual Scene Analysis (CAVSA). In CAVSA the scene is represented with audio and visual proto-objects. Proto-objects for the same speaker are grouped together in auditory and visual Short-Term Memories (STM) respectively. Visual and audio proto-objects are then matched based on position. Fig. 1 schematically illustrates the system architecture of CAVSA. Proto-objects, STM and audiovisual association are described below.

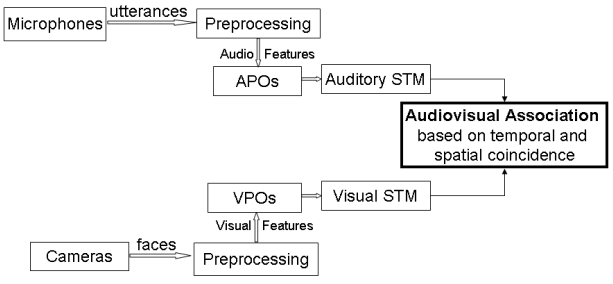

Fig. 1. System architecture of CAVSA. APO: audio proto-object, VPO: visual proto-object.

A. Proto-objects

Proto-objects are a psychophysical concept and are considered here as a compressed form of various features. Proto-objects can be tracked, pointed or referred to without identification and enable a flexible interface to behavior-control in robotics. For more information about proto-objects see [10], [11].

1) Visual proto-objects: In the camera field of view the visual objects can be segmented based on e.g. the similarity of color or shape to a given reference model. A frontal face detection algorithm based on [19] is used to extract visual proto-objects in multi-person scenarios. We assume that the person talking to robot is most of the time looking at the robot face. For each of these proto-objects, the center of the segment in the camera image is computed. The distance between robot head and speakers is about 1m. Using the distance information and saccade maps (see [20]), we can calculate the face position in 3D world coordinates and motor coordinates, and store them in the visual proto-object. Actually, proto-objects could contain an arbitrary number of features. Depending on the tasks, other features such as object size, orientation, color histogram and texture could also be used.

2) Audio proto-objects: A Gammatone Filterbank (GFB) as a model of the human cochlea is employed in the auditory preprocessing [11]. The GFB has 100 frequency channels that span the range of 100 -11000 Hz. To form audio proto-objects, we first segment audio streams based on energy. An audio segment begins when the signal energy exceeds a threshold and ends when the energy falls below this threshold. We then derive start time, length and energy for an audio proto-object. A filtering of audio proto-objects based on segment length and energy is then performed, since short or low power auditory signals are very probably noise. In addition, an audio proto-object contains also population-coded position cues (IID and ITD) and the estimated position encoded as a population vector, which will be explained in section III.

B. Short-term memory

In short-term memory (STM), proto-objects for the same speaker are grouped together. When a new proto-object appears, the procedure of entering it into STM can be described as follows:

- If the STM is empty, the new proto-object is added to the STM.
- If the STM already contains one or more proto-objects, the distance or similarity of selected grouping features are computed between the new proto-object and all proto-objects in the STM. If the distance between the new proto-object and the closest proto-object in the STM is smaller than a threshold, these two proto-objects are merged (averaged). Otherwise the new proto-object is inserted into the STM.
- Proto-objects, which are not updated for more than a certain period (200 s in our experiment) are removed from the STM.

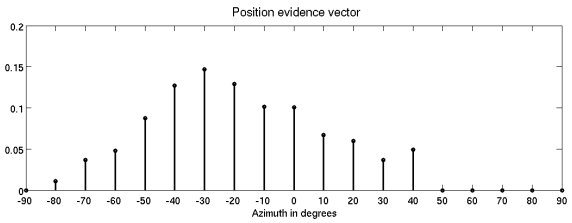

Fig. 2. An example of a position evidence vector, corresponding to output firing rates from a bank of neurons with different receptive fields. Here the estimated azimuth is $-30°$.

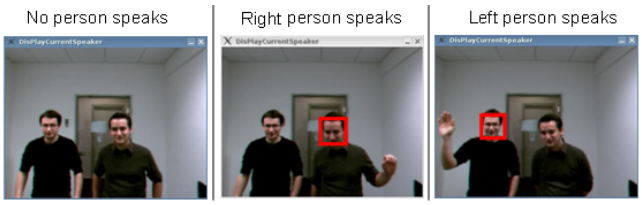

Fig. 3. Detection of the current speaker. Face detection of the current speaker is visualized. To easily evaluate the results, the current speaker was required to raise one hand.

Using such a STM, it is not necessary to buffer all the incoming proto-objects for processing. Moreover, we can match an audio proto-object to a visual proto-object, even if it is out of sight for a while e.g. due to the movements of robot head.

In auditory STM, grouping features could be position and/or spectral energies. In visual STM one or more features among position, color and size could be employed to group visual proto-objects. In this work only position is used as a grouping feature for the auditory and visual STM. In an audio proto-object the position is represented by population code vector. For more information about this population coding see section III. The similarity of position vectors between the new audio proto-object and an audio proto-object in auditory STM is based on the scalar product of normalized position vectors (mean 0, norm 1). We set the threshold of the similarity to be 0.6 empirically. In visual STM the Euclidean distance of positions in 3D world coordinates is calculated and the threshold is set to 30 cm, so that slight movements of speakers such as head shaking are tolerated.

C. Audiovisual association

Position is also used to match a visual proto object and an audio proto-object from their STMs. Auditory position evidence vectors and visual positions in world coordinates must be converted to the same metric, for which motor coordinates (azimuth and elevation) are preferred. Speakers are about 1m away from the robot. In this paper we concentrate only on azimuth as mentioned. The azimuth of an audio proto-object is taken as the peak position in its position evidence vector, while the azimuth of a visual proto-object is estimated using saccade maps (see [20]). Fig. 2 shows an example of a position evidence vector.

The azimuth distance between audio proto-object A_i $(i \in [1,M])$ and visual proto-object V_j $(j \in [1,N])$ is denoted as $\Delta X(A_i,V_j)$, where M and N stand for the number of audio and visual proto-objects in auditory and visual STM respectively. The relative probability that audio proto-object A_i belongs to visual proto-object V_j can then be approximated as:

$$P_{common}(A_i,V_j) \approx \exp\left(\frac{-\Delta X(A_i,V_j)^2}{2\cdot\delta_{AV}^2}\right), \qquad (1)$$

where the standard deviation δ_{AV} represents the average difference in estimated azimuth between an audio and a visual proto-object which are caused by the same speaker. Next, we check how certain the association between visual proto-object V_{jMax} with the maximal probability and A_i is. If one visual proto-object shows a very high probability and all other visual proto-objects have a low probability, this indicates a reliable association. Conversely, when all visual proto-objects have quasi equal probability, the association is unreliable. The uncertainty is computed using the entropy of the normalized probability. A similar usage of entropy can be found in speech recognition, as per [21]. All probabilities $P_{common}(A_i,V_j)$ for audio proto-object A_i are normalized such that they sum up to 1. The normalized probability is denoted as:

$$\widehat{P}_{common}(A_i,V_j) = \frac{P_{common}(A_i,V_j)}{\sum_{j=1}^{N} P_{common}(A_i,V_j)}. \qquad (2)$$

The entropy for A_i is given as:

$$H_i = \begin{cases} 0 & \text{if } N=1, \\ \dfrac{-\sum_{j=1}^{N}\left(\widehat{P}_{common}(A_i,V_j)\cdot\log_2\widehat{P}_{common}(A_i,V_j)\right)}{\log_2 N} & \text{if } N>1. \end{cases} \qquad (3)$$

Here, the division by $\log_2 N$ ensures that the maximal H_i is 1 to easily set a threshold Θ_H. If entropy H_i is larger than Θ_H, A_i and V_{jMax} are not associated. Θ_H was set to 0.8 empirically. The uncertainty of the whole audiovisual association can be captured by averaging over all H_i :

$$H = \frac{\sum_{i=1}^{M} H_i}{M}. \qquad (4)$$

Given learned audio-motor maps, Fig. 3 illustrates an example, where CAVSA is used in an online scenario to visualize the current active speaker among two speakers who stand in front of the robot.

III. Online Adaptation

In this section we will describe how to find the matched visual position to the current sound and how to adapt audio-motor maps. In our system audio-motor maps represent the relation between population-coded cues and position evidence vectors. An audio-motor map M contains for each azimuth angle p (-90, -80, ..., 0, ..., 80, 90), each cue l ($l=1$ for IID, $l=2$ for ITD) and each frequency channel f $(1-100)$ a population code vector $M(p,l,f,n)$. Nodes (neurons) n have receptive field centers at $(-0.9,-0.8,...,0,...,0.8,0.9)$. We measure binaural cues in each frequency channel when

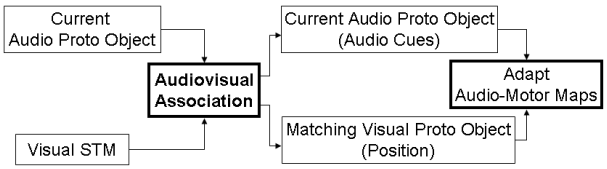

Fig. 4. System architecture of online adaptation using CAVSA

an onset appears, encode then the measured cues and store them in the audio proto-object. For encoding, the same set of neurons n is used and every measured cue IID or ITD leads to an activation in the nearest neurons, so that a population code vector is generated. In each frequency channel f, population code vectors for all measurements of cue l are then summed up in an audio proto-object. Finally, each summed population code vector is normalized to mean 0 and norm 1. Let us denote the encoded cue l in frequency channel f, at node n as $C(l,f,n)$. To acquire position evidence vector $E(p)$, population response $C(l,f,n)$ is compared with stored population responses $M(p,l,f,n)$ for all positions p by computing scalar products. The peak in position vector $E(p)$ is taken as the estimated sound source position. For more details see [11].

We use only the current audio proto-object instead of the auditory STM ($M = 1$), since only information about the current sound is required. The number of proto-objects in visual STM (N) is considered as the number of speakers in the dialog scenario. The system architecture of online adaptation using CAVSA is illustrated in Fig. 4.

The matched visual proto-object is searched for using equation (1), (2) and (3). Note that if $N = 1$, then $\widehat{P}_{common}(A,V) = 1$ and entropy $H = 0$. That means, if only one visual proto-object exists in the visual STM, the audio and visual proto-object are assumed to have a common cause. During the learning of the audio-motor maps, the standard deviation δ_{AV} in equation (1) is dynamically updated depending on the quality of the current audio-motor map. We approximate δ_{AV} by calculating the average difference between estimated position in audio and visual proto-objects over time using the following update rule:

$$\delta_{AV}^{t} = \begin{cases} \Delta X(A,V) \cdot w + \delta_{AV}^{t-1} \cdot (1-w) & \text{if } N = 1, \\ \delta_{AV}^{t-1} & \text{otherwise.} \end{cases} \quad (5)$$

Here, t and w represent update step and weight respectively. We set $w = 0.1 \cdot \beta$ dependent on the fixed adaptation rate β, which controls the degree of adaptation for a single step. $\Delta X(A,V)$ describes the position distance between the audio and visual proto-object in the current adaptation step. δ_{AV}^{t} is updated only if just one visual proto-object exists. The initial value δ_{AV}^{0} is set to 40° empirically.

In the experiments it was found that a visual proto-object, which is not related to the current sound source but near the correct visual proto object, can also enhance the quality of audio-motor maps, particularly when the quality of maps is poor as during initialization. Thus if entropy H exceeds the threshold Θ_H, but the position distance between the visual proto objects with maximum and second maximum probability ($\widehat{P}_{common}$) is small, audio-motor maps can be updated nonetheless. The uncertainty of an adaptation step can be described by the following equation:

$$H' = H \cdot \Delta X(V_{jMax}, V_{jSecMax}), \quad (6)$$

where V_{jMax} and $V_{jSecMax}$ stand for the visual proto-objects with maximal and second maximal probability respectively. If uncertainty H' is below threshold $\Theta_{H'}$ or $H < \Theta_H$, a confidence factor c is set to 1 and the map is adapted. Otherwise $c = 0$ and the map is not updated in the current step. The threshold $\Theta_{H'}$ depends on the standard deviation δ_{AV} ($\Theta_{H'} = 2 \cdot \delta_{AV}$), since the system has a high tolerance for the visual position difference when the quality of audio-motor maps is poor. The matched visual position p_v is then converted to a position evidence vector, which can be defined by a delta function δ_{p,p_v}.
The audio-motor map is updated by:

$$\begin{aligned} M(p,l,f,n,t) \quad = \quad & M(p,l,f,n,t-1) - F(p) \cdot \\ & (M(p,l,f,n,t-1) - C(l,f,n)), \end{aligned} \quad (7)$$

where p, l, f and n stand for position, cue index, frequency channel and node, respectively. Learning parameter $F(p)$ is given by:

$$F(p) = c \cdot \beta \cdot \delta_{p,p_v}, \quad (8)$$

where c and β represent the confidence of the matching process and the fixed adaptation rate respectively. In our experiment $\beta = 0.2$.

IV. Results

Our approach was tested in real world scenarios. Offline-calibrated maps were used as reference. Our approach was compared with a heuristic method in scenarios where additional persons dynamically entered and vacated the room.

A. Offline-calibrated audio-motor maps as reference

In the experiment we firstly calibrated audio-motor maps offline und used them as reference for performance estimation. A loudspeaker was placed in front of the robot (0°), at a distance of 1m away and at the same height as the robot head. The head changed its orientation p_h every 10° from -90° to 90°, so that the azimuth ($-p_h$) changed correspondingly in robot-centered coordinates. At each position, 47 sound files were played and mean population responses of IID and ITD were measured. The whole offline calibration required more than 2 hours.

The performance of online-adapted audio-motor maps can be then estimated by comparison with offline-calibrated maps using normalized Euclidean distance:

$$d(M,M') = \sqrt{\frac{\sum_p \sum_l \sum_f \sum_n (M(p,l,f,n) - M'(p,l,f,n))^2}{K}}, \quad (9)$$

where M and M' represent online-adapted and offline-calibrated maps respectively. K is the total number of elements in an audio-motor map and satisfies $K = k_p \cdot k_l \cdot k_f \cdot k_n$, where $k_p = 19$, $k_l = 2$, $k_f = 100$ and $k_n = 19$ is the number of positions, cues, frequency channels and nodes, respectively.

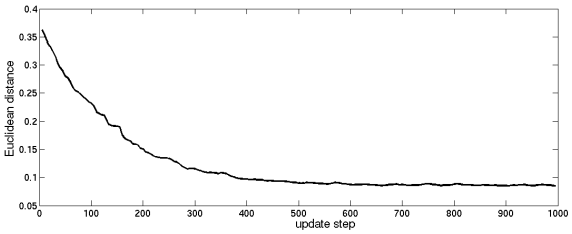

Fig. 5. Euclidean distance between offline-calibrated and online-adapted maps over time

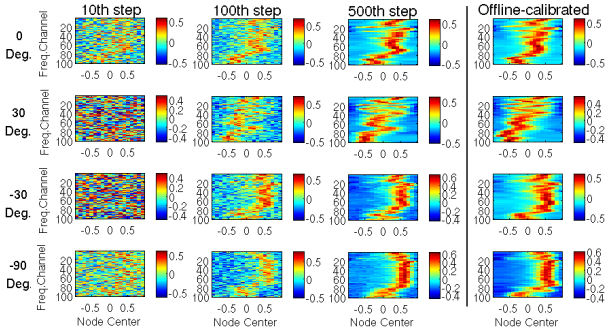

Fig. 6. Online-adapted IID maps for several azimuth angles and in different adaptation steps. Offline-calibrated maps are used as reference.

B. Basic online scenario

In the online scenario we simulated a speaker with a loudspeaker on which a picture of a face was attached. The loudspeaker was placed on the same position as in offline calibration. During online adaptation, the robot head oriented itself to a random horizontal angle in the range $[-90, 90]$ after an update step was finished. The acquisition of auditory and visual signals was interrupted during head movement, so that audio-motor maps were only adapted in still status. At the beginning maps are initialized with random numbers in the range $[-0.5, 0.5]$ using a uniform distribution. The normalized Euclidean distance over time between online-adapted and offline-calibrated maps is illustrated in Fig. 5. Fig. 6 shows online-adapted IID maps on several positions ($0°$, $30°$, $-30°$ and $-90°$) and in different update steps (10th, 100th and 500th step), as well as offline-calibrated maps on the corresponding positions. Every 100 update steps of our approach need about 7 minutes. The plot in Fig. 5 shows that a good similarity is achieved after about 400 steps, which takes about 30 minutes, while offline calibration requires more than 2 hours.

C. Natural communication

Our approach was tested in three scenarios where additional persons ($N > 1$) dynamically entered and vacated the scene. The results were then compared with a heuristic method 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. The heuristic method is similar to methods in [8], [9] for linking auditory and visual information. The three scenarios and the corresponding results are

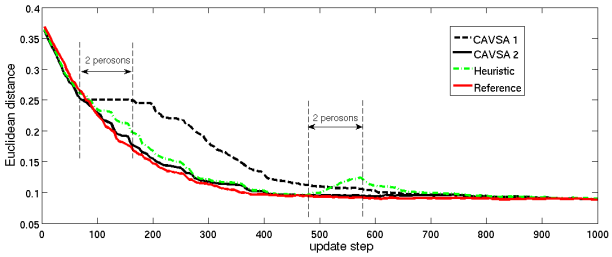

Fig. 7. Scenario 1: one additional person entered the room in the 70th adaptation step and vacated in the 170th adaptation step. He entered then in the 480th step and vacated in the 580th step again.

described as follows.

The difference between the first scenario and the basic online scenario in section IV-B is that an additional person entered the room during the online adaptation in the first scenario, stood 1m away, faced the robot for a while, did not speak and then vacated. After some update steps the person entered the room and vacated again in the same manner. The only sound source was the loudspeaker at $0°$, since the additional person did not speak. For this scenario, CAVSA which computes the association uncertainty with H in equation (3), CAVSA with consideration of both H and H' in equation (6), the heuristic method and the method using known sound source position ($0°$) as reference are compared. To simplify the description let us denote these four methods as "CAVSA 1", "CAVSA 2", "Heuristic" and "Reference", respectively. As shown in Fig. 7, the quality of audio-motor maps was still poor between the 70th and 170th adaptation step, when the additional person was in the room for the first time. "CAVSA 1" did not update the maps due to the high entropy H in equation (3). "Heuristic" selected sometimes the wrong position for adaptation, but improved the performance of audio-motor maps to some degree because the quality of the maps was still poor. "CAVSA 2" nearly reached the performance of "Reference" which used true sound source position. Between the 480th and 580th step the maps were refined. "CAVSA 1" and "CAVSA 2" were almost not influenced by the additional person because of their good performance on audiovisual integration, while the error in "Heuristic" increased due to using wrong positions.

The only difference between the first and second scenario is that two additional persons instead of one dynamically entered the room. Fig. 8 shows the comparison of the four methods with Euclidean distance to offline-calibrated maps over time. In comparison to Fig. 7 the four methods performed similarly except that "Heuristic" got much worse when two additional persons appeared than in the first scenario when only one additional person appeared.

In the third scenario the loudspeaker was not used. Instead two speakers stood 1m away from the robot, faced the robot and talked to it alternatingly. After some steps one person vacated the room and only one spoke to the robot. The adaptation began with an audio-motor map which had been adapted for 80 steps. Fig. 9 illustrates the comparison

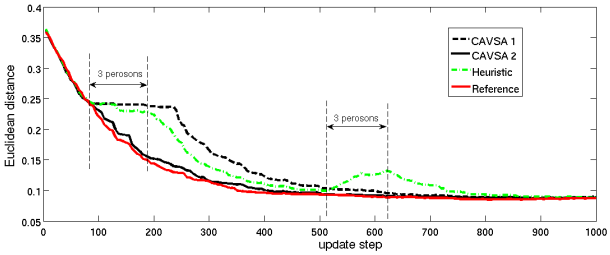

Fig. 8. Scenario 2: two additional persons entered the room in the 80th adaptation step and vacated in the 190th step. They entered then in the 515th step and vacated in the 620th step again.

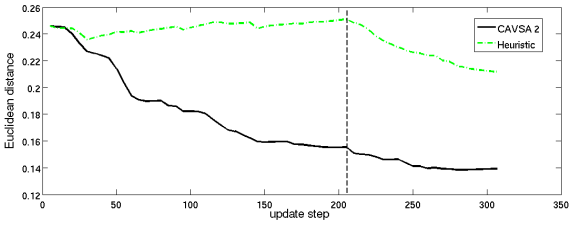

Fig. 9. Scenario 3: from start two speakers talked to the robot alternatingly. In the 205th step one person vacated the room and only one spoke to the robot.

of "CAVSA 2" and "Heuristic" up to the 310th adaptation step. It was shown that "CAVSA 2" performed much better than "Heuristic" when two speakers talked to the robot alternatingly.

The results in these three scenarios showed that the adaptation process with CAVSA was more robust in situations where additional persons dynamically entered and vacated the scene.

V. Summary and Outlook

We have suggested an approach for Computational AudioVisual Scene Analysis (CAVSA) with a focus on human-robot interaction in multi-person environments. 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 on position information. We have shown that our system can correctly determine the number and position of speakers in typically human-robot dialog scenarios. This was demonstrated by the online adaptation of audio-motor maps. Comparing our online adaptation of audio-motor maps using CAVSA with prior online adaptation methods, our approach is more robust in situations with more than one speaker and when speakers dynamically enter and leave the scene. Only spatial coincidence is so far used to group audio and visual proto-objects in their STMs, which fails for instance when a person moves quickly or several persons stand very close to each other. Hence we plan to employ more grouping features such as spectral energies for auditory STM and color or size for visual STM.

VI. Acknowledgements

Sincere thanks to Bram Bolder and Stefan Kirstein for much helpful support with the vision system. This work is supported by the Honda Research Institute Europe.

References

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

[2] H. Finger, P. Ruvolo, S. Liu, and J. R. Movellan, "Approaches and databases for online calibration of binaural sound localization for robotic heads," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, October 2010.

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

[4] M. M. V. Wanrooij and A. J. V. Opstal, "Relearning sound localization with a new ear," *Neuroscience*, vol. 25, pp. 5413–5424, June 2005.

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

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

[7] ——, "Early blindness results in a degraded auditory map of space in the optic tectum of the barn owl," *Proc Natl Acad Sci USA*, vol. 85, pp. 6211–6214, 1998.

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

[9] H. Nakashima and N. Ohnishi, "Acquiring localization ability by interaction between motion and sensing," *IEEE International Conference on Systems, Man and Cybernetics*, October 1999.

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

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

[12] U. V. Chaudhari, G. N. Ramaswamy, G. Potamianos, and C. Neti, "Audio-visual speaker recognition using time-varying stream reliability prediction," *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. vol. V, pp. 712–715, April 2003.

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

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

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

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

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

[18] H. Liu and M. Shen, "Continuous sound source localization based on microphone array for mobile robots," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, October 2010.

[19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[20] T. Rodemann, F. Joublin, and C. Goerick, "Continuous and robust saccade adaptation in a real-world environment," *KI-Kuenstliche Intelligenz*, March 2006.

[21] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, January 2002.