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

Learning of Audiovisual Integration

Rujiao Yan, Tobias Rodemann, Britta Wrede

2011

Preprint:

This is an accepted article published in Proceedings of ICDL/EPIROB 2011. The final authenticated version is available online at: https://doi.org/[DOI not available]

Learning of Audiovisual Integration

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

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

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

***Abstract*—We present a system for learning audiovisual integration based on temporal and spatial coincidence. The current sound is sometimes related to a visual signal that has not yet been seen, we consider this situation as well. Our learning algorithm is tested in online adaptation of audio-motor maps. Since audio-motor maps are not reliable at the beginning of the experiment, learning is bootstrapped using temporal coincidence when there is only one auditory and one visual stimulus. In the course of time, the system can automatically decide to use both spatial and temporal coincidence depending on the quality of maps and the number of visual sources. We can show that this audiovisual integration can work when more than one visual source appears. The integration performance does not decrease when the related visual source has not yet been spotted. The experiment is executed on a humanoid robot head.**

I. Introduction

The integration of auditory and visual information derived from the same event can enhance the representation of the external world. Therefore, audiovisual integration is used in a broad range of applications, such as speaker recognition and speaker tracking in robotics, as well as speaker indexing in multimedia data [1]–[3]. It is known that the brain performs well in integrating related information from audition and vision. Hence it makes sense to understand the way how the brain integrates auditory and visual stimuli, and develop an algorithm to describe the behavior.

There is evidence in support of the view that the connection between human vision and audition is present already to some degree at birth, and the integration ability is then developed experience-dependent [4]. Temporal coincidence has been identified as one of the most important factors determining whether or not multisensory integration takes place [5]. 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 [6]. If more than one source exists, we need other information such as position information to avoid ambiguity. The closer a visual stimulus is to an auditory stimulus, the more probable they are perceived as having a common cause [7]. 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 [8], [9]. When the animals are raised in artificial environments where auditory and visual stimuli are temporally coupled but spatially not coherent, multisensory neurons in the superior colliculus (SC) are also able to integrate these stimuli. Spatial coincidence appears to be learned early in life adaptive to the environment of an animal to deal with that environment well later in life.

Hershey et al. [10] use only temporal coincidence between lip motion and speech for audiovisual integration. The approach works when the sound sources are always in the view. Our learning algorithm is based on both temporal and spatial coincidence, and we consider the situation where the current sound is related to a visual signal that has not yet been seen. The algorithm is tested in online adaptation of audio-motor maps. Audio-motor maps describe the relationship between audio cues and sound position in motor coordinates (azimuth and elevation). These audio cues such as interaural time difference (ITD) and interaural intensity difference (IID), result from the interaction of the head and ears with the incoming auditory stimulus [11]. Using audio-motor maps we can obtain sound source positions from measured audio cues. Since vision plays an important role in calibration of audio-motor maps in humans and animals [12]–[14], it is used as the feedback signal for precise position information. It is then necessary to match a visual signal to the current sound using audiovisual integration, 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, such as in the ventriloquism aftereffect [15]. Given precise measurements of visual position and audio cues, the quality of maps depends on the performance of audiovisual integration.

Natale et al. [16] use temporal coincidence between motion and sound to integrate auditory and visual stimuli for learning both saccade maps and audio-motor maps. Their regime works only under the assumption that no object except the current sound source moves in the view. Other methods adapt audio-motor maps using visual feedback [17], [18]. The approach in [17] fails when more than one visual source or an unrelated visual source appears. Nakashima et al. [18] attempt the online adaptation of audio-motor maps in a simplified environment, where a red marker is attached to the sound source and no other red object exists. In comparison to these methods, we intend to adapt audio-motor-maps in more complex environments using audiovisual integration.

Since audio-motor maps are not reliable at the beginning of the experiment, learning is bootstrapped using temporal

coincidence when there is only one auditory and one visual stimulus. This may be seen as analogous to biology where many animals already have audio-motor maps at birth, but the maps are very rough and need to be calibrated by experience [13]. Similarly, our system can automatically decide to use both spatial and temporal coincidence in the course of time depending on the quality of maps and the number of visual sources.

II. Audiovisual integration

In this section we introduce audiovisual integration in scenarios where the current acoustic signal is to be related to one out of many visual signals, for instance we hear a sound and see many faces. Position is used as correlation information between auditory and visual signals. The difference between the position of the current sound source p_a and the position of a visual signal p_{v_i} $(i \in [1,N])$ is denoted as $d(p_a, p_{v_i})$, where N stands for the number of visual signals. Then the relative probability that the visual signal at position p_{v_i} belongs to the current auditory signal is approximated by a Gaussian function:

$$P_{common}(p_a, p_{v_i}) = \exp\left(-\frac{d(p_a, p_{v_i})^2}{2 \cdot \delta_{AV}^2}\right), \tag{1}$$

where the standard deviation δ_{AV} represents the average difference in estimated position between an auditory and a visual signal which are caused by the same physical object. Next, we check the entropy of the set of normalized probabilities to confirm that the maximal probability P_{common} is valid. If one visual signal shows a very high probability and all other visual signals have low probabilities, this expresses a low entropy indicating a reliable integration. Conversely, when all visual signals have quasi equal probability, the entropy is high and the integration is unreliable. We calculate the entropy of normalized probabilities in a manner similar to that found in speech recognition [19]. All probabilities $P_{common}(p_a, p_{v_i})$ are normalized such that they sum to 1, and the normalized probability is denoted as $\widehat{P}_{common}(p_a, p_{v_i})$:

$$\widehat{P}_{common}(p_a, p_{v_i}) = \frac{P_{common}(p_a, p_{v_i})}{\sum_{i=1}^{N} P_{common}(p_a, p_{v_i})}. \tag{2}$$

The entropy is then computed as follows:

$$H = \begin{cases} 0 & \text{if } N = 1, \\ \frac{\sum_{i=1}^{N} \widehat{P}_{common}(p_a, p_{v_i}) \cdot \log_2 \widehat{P}_{common}(p_a, p_{v_i})}{\log_2 N} & \text{if } N > 1, \end{cases} \tag{3}$$

where the division by $\log_2 N$ ensures that the maximal value of H is 1. If entropy H is larger than a threshold Θ_H, the current auditory signal is not linked to any visual signal. Θ_H is set to 0.8 empirically.

Now we take account of the situation where the current auditory signal is related to a visual signal that has not yet been seen. For convenience, we denote the position of the unseen visual signal as $p_{v_{N+1}}$. The probability that the current auditory signal is associated with the visual signal at position $p_{v_{N+1}}$ is described by:

$$P_{common}(p_a, p_{v_{N+1}}) = \frac{\sum_{p} U(p) \cdot P_{common}(p_a, p)}{\sum_{p} P_{common}(p_a, p)}. \tag{4}$$

Here, $U(p)$ stands for a view memory with elements between 0 and 1 for each position p. The larger the value $U(p)$ at p is, the longer the time since position p has not been attended. $U(p) = 0$ means that position p is currently in the field of view. The normalized probability and entropy become:

$$\widehat{P}_{common}(p_a, p_{v_i}) = \frac{P_{common}(p_a, p_{v_i})}{\sum_{i=1}^{N+1} P_{common}(p_a, p_{v_i})}, \tag{5}$$

and

$$H = \frac{\sum_{i=1}^{N+1} \widehat{P}_{common}(p_a, p_{v_i}) \cdot \log_2 \widehat{P}_{common}(p_a, p_{v_i})}{\log_2 (N+1)}. \tag{6}$$

Note that if $P_{common}(p_a, p_{v_{N+1}})$ is larger than the maximal $P_{common}(p_a, p_{v_i})$ $(i \in [1,N])$, the matched visual signal is not seen, and the audiovisual integration can not be executed. Furthermore, if $N = 1$ and $P_{common}(p_a, p_{v_{N+1}}) \to 0$, 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. In this situation, only temporal coincidence is employed for audiovisual integration.

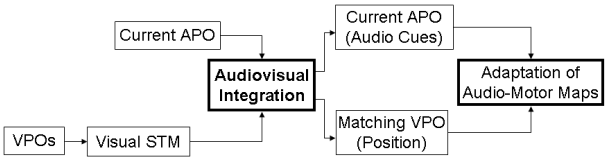

Fig. 1. System architecture of online adaptation using audiovisual integration. APO: audio proto-object, VPO: visual proto-object.

III. Test in online adaptation of audio-motor maps

The presented algorithm of audiovisual integration is employed in online-adaptation of audio-motor maps to find the visual position matching the current sound. Firstly, auditory and visual signals are represented in form of proto-objects. The concept of proto-object is explained in Section III-B. Visual proto-objects for the common origin (the same speaker) are grouped together in short-term memory (STM). Then the audiovisual integration method described in Section II is employed to find the matched visual proto-object. Finally, audio-motor maps are adapted using the matched visual position. Fig. 1 schematically illustrates the system architecture of online adaptation using audiovisual integration. The experiment is conducted using a humanoid robot head with a pair of cameras and a pair of microphones. The head is mounted on a pan-tilt unit, and just the left camera is employed to capture the visual signal. Moreover, a Gammatone Filterbank (GFB) is

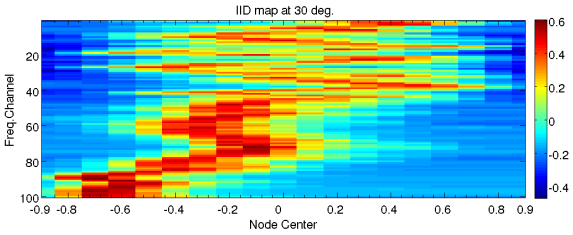

Fig. 2. An example of the IID map at 30°.

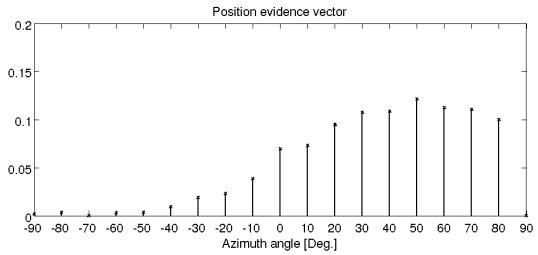

Fig. 3. An example of a position evidence vector, where the estimated azimuth angle is 50°.

employed in the auditory preprocessing [20]. The GFB has 100 frequency channels that span the range of 100 -11000 Hz.

A. Audio-motor map

In this work an audio-motor map represents the relationship between population-coded cues and position evidence vectors. To ease the description, only the azimuth is considered. Azimuth positions between $-90°$ and $90°$ are taken into account because of the mechanical constraint of the robot head. An audio-motor map is denoted as M which 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 n have response centers at $(-0.9,-0.8,...,0,...,0.8,0.9)$. Fig. 2 illustrates an example of an IID map $M(30,1,f,n)$ at 30°. For more information on audio-motor maps see [20].

B. Visual and auditory representation

Various auditory and visual features are collected in audio and visual proto-objects respectively. A proto-object is a psychophysical concept and is considered here as a compressed form of a set of features. A proto-object can be tracked, pointed or referred to without identification. For more information on proto-objects see [20], [21].

In the camera field of view we use a face detection algorithm based on [22] to extract visual proto-objects. For each of these proto-objects, the center of the segment in the camera image is computed. Participants are placed approximately 1m away from the robot. Within one visual proto-object we store the position of the face in camera image and in 3D world coordinates. Next, visual proto-objects for the common origin are grouped together in short-term memory (STM). When a new proto-object appears, the procedure of entering it into STM can be described as follows:

1) If the STM is empty, the new proto-object is added to the STM.
2) 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.
3) Proto-objects that are not updated for more than a certain period ($T = 100s$ in our experiment) are removed from the STM.

Using such a STM it is not necessary to store all the incoming visual proto-objects for processing, for instance when the same face appears in different image frames. Moreover, we can match the current audio proto-object to a visual proto-object, even if it is out of sight for some time. The number of visual proto-objects in STM, N, is described in Section II.

Object position in 3D world coordinates is used as the grouping feature. Euclidean distance of positions between the new proto-object and each proto-object in the STM is calculated and the threshold is set to 30 cm, so that slight movements of participants such as head shaking are tolerated.

To form audio proto-objects, we first segment audio streams based on energy. An audio proto-object begins when the signal energy exceeds a threshold and ends when the energy falls below this threshold. Since short or low power auditory signals are very probably noise, a filtering of audio proto-objects based on segment length and energy is performed, as per [20]. In our experiment an audio proto-object contains a start time, segment length, energy of a segment, population-coded cues (IID and ITD) and a position evidence vector. For encoding of audio cues, the same set of nodes n as in audio-motor map M is used and every measured cue IID or ITD leads to an activation in the nearest nodes. All measurements are added over time for an audio proto-object. For each frequency channel in encoded cues, the population code vector is normalized to mean 0 and norm 1. Let us denote the population-coded 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 audio-motor maps $M(p,l,f,n)$ for all positions p by computing scalar products. The peak in position vector $E(p)$ is taken as the estimated sound source position. Fig. 3 shows an example of a position evidence vector.

C. Audiovisual integration in online adaptation

Position is used as correlation information for audiovisual integration. Thus auditory position evidence vectors and visual positions in world coordinates must be converted to the same metric, for which motor coordinates are preferred. Participants are placed about 1m away from the robot. 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

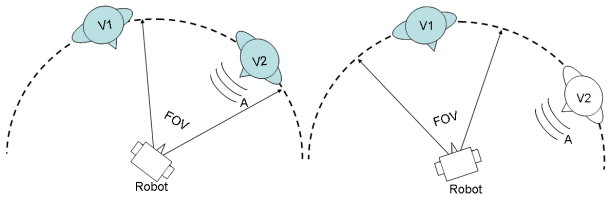

Fig. 4. Examples: *left*: both visual sources have been seen; *right*: the matched visual proto-object V_2 has not yet been seen. A is the current audio proto-object, V_1 and V_2 are visual proto-objects. The field of view (FOV) is 60°.

proto-object is estimated using saccade maps (see [23], [24]). We denote the current audio proto-object by A and a visual proto-object in the STM as V_i ($i \in [1,N]$). In the following subsections, we first introduce two methods of audiovisual integration. One (see Algorithm 2) considers the situation where the matched visual proto-object is not in the STM, while the other (see Algorithm 1) does not. The performance of these two methods will be compared in Section IV. We then take account of the uncertainty of an adaptation step to accelerate the online adaptation process.

1) Basic approach: Assuming all visual sources have been seen as the example shown in Fig. 4 (*left*), we compute the uncertainty of the current audiovisual integration using entropy H in Eq. (3). During the learning of the audio-motor maps, the standard deviation δ_{AV} as per Eq. (1) is dynamically updated depending on the quality of the current audio-motor map. We obtain δ_{AV} by calculating the average difference between estimated azimuth angle in audio and visual proto-objects over time using the following update rule:

$$\delta_{AV}^{s} = \begin{cases} d(p_a, p_v) \cdot w + \delta_{AV}^{s-1} \cdot (1-w) & \text{if } N = 1, \\ \delta_{AV}^{s-1} & \text{otherwise.} \end{cases} \quad (7)$$

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. (13)). The difference of azimuth angles between the audio and the visual proto-object in the current adaptation step is denoted as $d(p_a, p_v)$. If only one visual proto-object is in the STM, δ_{AV}^{s} is updated. The basic approach is also described in Algorithm 1.

2) Consideration of unseen visual proto-objects: We consider now the situation where the current audio proto-object A is related to visual proto-object V_{N+1} that is not stored in the STM, as the example shown in Fig. 4 (*right*). The process is described in Algorithm 2. The probability that A and V_{N+1} are caused by the same speaker is calculated using Eq. (4). The view memory $U(p)$ in Eq. (4) is computed with $U(p) = 1 - a(p,t^v)$. For each azimuth angle p, activity function $a(p,t^v)$ is defined as:

$$a(p,t^v) = e^{-\frac{t^v}{T}}, \quad (8)$$

where t^v represents the time in seconds since position p has been viewed the last time. The initial value of t^v is set to ∞, so that $a(p,t^v) = 0$ if position p has never been attended. Parameter $T = 100s$ is used to decay the activity.

Algorithm 1 Audiovisual integration: Basic approach

for $i = 1$ to N **do**
  Calculate $P_{common}(p_a, p_{v_i})$ based on Eq. (1)
end for
for $i = 1$ to N **do**
  Calculate $\widehat{P}_{common}(p_a, p_{v_i})$ based on Eq. (2)
end for
Calculate entropy H as in Eq. (3)
for $i = 1$ to N **do**
  Search for the visual proto-object V_{Max} that has the maximal $P_{common}(p_a, p_{v_i})$
end for
if $H < \Theta_H$ **then**
  Integrate A with V_{Max}
end if
if $N = 1$ **then**
  Update standard deviation δ_{AV} as in Eq. (7)
end if

Standard deviation δ_{AV} in Eq. (1) is updated only if just one visual proto-object exists and positions near the current proto-object have been recently visually attended. The update rule is described as below:

$$\delta_{AV}^{s} = \begin{cases} d(p_a, p_v) \cdot w + \delta_{AV}^{s-1} \cdot (1-w) & \text{if } N = 1 \text{ AND } P_u < \Theta_{P_u}, \\ \delta_{AV}^{s-1} & \text{otherwise.} \end{cases} \quad (9)$$

Here, $P_u = \widehat{P}_{common}(p_a, p_{v_{N+1}})$ is the normalized probability that A and V_{N+1} have a common cause, as described in Eq. (5). Θ_{P_u} is the threshold of P_u and is set to 0.1.

Algorithm 2 Audiovisual integration: Consideration of unseen visual proto-objects

for $i = 1$ to N **do**
  Calculate $P_{common}(p_a, p_{v_i})$ based on Eq. (1)
end for
Calculate $P_{common}(p_a, p_{v_{N+1}})$, the probability that A and V_{N+1} have a common cause, as in Eq. (4)
for $i = 1$ to $N+1$ **do**
  Calculate $\widehat{P}_{common}(p_a, p_{v_i})$ based on Eq. (5)
end for
Calculate entropy H as in Eq. (6)
for $i = 1$ to $N+1$ **do**
  Search for the visual proto-object V_{Max} that has the maximal $P_{common}(p_a, p_{v_i})$
end for
if $H < \Theta_H$ AND $V_{Max} \neq V_{N+1}$ **then**
  Integrate A with V_{Max}
end if
if $N = 1$ AND $P_u < \Theta_{P_u}$ **then**
  Update standard deviation δ_{AV} as in Eq. (9)
end if

3) Uncertainty of an adaptation step: Comparing entropy H with threshold Θ_H, we can decide whether the current audio proto-object A and the visual proto-object with the maximal probability (P_{common}) are integrated. However, it is found in the experiments that a candidate visual proto-object, which is not related to the current sound source but which is positioned 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:

$$H' = H \cdot d(p_{v_1}, p_{v2}), \tag{10}$$

where p_{v_1} and p_{v2} stand for the positions of 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 confidence factor c is given by:

$$c = \begin{cases} 1 & \text{if } H < \Theta_H \text{ OR } H' < \Theta_{H'}, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

In this manner the adaptation process is accelerated. 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.

D. Online adaptation of audio-motor maps

By means of population-coded cues $C(l,f,n)$ in the current audio proto-object and the matched visual position $p_{v'}$, audio-motor map M is updated by:

$$\begin{aligned} M^s(p,l,f,n) &= M^{s-1}(p,l,f,n) - F(p) \cdot \\ &\quad (M^{s-1}(p,l,f,n) - C(l,f,n)), \end{aligned} \tag{12}$$

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

$$F(p) = c \cdot \beta \cdot \delta_{p,p_{v'}}, \tag{13}$$

where c and β represent the confidence of the matched process and the fixed adaptation rate respectively. In our experiment $\beta = 0.2$. Position evidence vector $\delta_{p,p_{v'}}$ is defined by a delta function:

$$\delta_{p,p_{v'}} = \begin{cases} 1 & \text{if } p = p_{v'}, \\ 0 & \text{if } p \neq p_{v'}. \end{cases} \tag{14}$$

Here, $p_{v'}$ represents the matched visual position. The online adaptation algorithm is described in Algorithm 3.

Algorithm 3 Online adaptation of audio-motor maps

```
Given C(l,f,n) in the current audio proto-object and p_v' in the matched visual proto-object
if H < Θ_H OR H' < Θ_H' then
    c ← 1
else
    c ← 0
end if
for p = −90 to 90 step 10 do
    if p = p_v' then
        δ_{p,p_v'} ← 1
    else
        δ_{p,p_v'} ← 0
    end if
    Calculate learning parameter F(p) based on Eq. (13)
end for
for p = −90 to 90 step 10 do
    for l = 1 to 2 step 1 do
        for f = 1 to 100 step 1 do
            for n = −0.9 to 0.9 step 0.1 do
                Update M^s(p,l,f,n) based on Eq. (12)
            end for
        end for
    end for
end for
```

IV. Results

Our approach was tested in real world scenarios where participants dynamically entered and vacated the room. Offline-calibrated maps were used as reference.

A. Offline-calibrated audio-motor maps as reference

We first calibrated audio-motor maps offline and used them as reference for performance estimation. For the calibration, a loudspeaker was placed in front of the robot (0°), at a distance of 1m and at the same height, as shown in Fig. 5. The head changed its orientation p_h every 10° from −90° to 90°, so that the azimuth angle of the loudspeaker ($-p_h$) changed correspondingly in robot-centered coordinates. At each position, 47 sound files were played and mean population responses of IID and ITD were measured. A similar offline-calibration approach, as per [25], uses ground truth positions provided by a motion capture system. The performance of online-adapted audio-motor maps can be estimated by comparison with offline-calibrated maps using normalized Euclidean distance:

$$d(M,M') = \sqrt{\frac{\sum_p \sum_l \sum_f \sum_n (M(p,l,f,n) - M'(p,l,f,n))^2}{K}}, \tag{15}$$

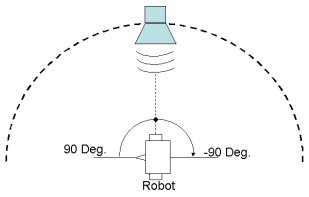

Fig. 5. Sketch of the experimental setting: offline calibration.

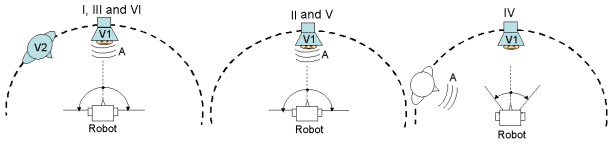

Fig. 6. Scenario 1: Sketch of the experimental setting in temporal phases I-VI. A is the current audio proto-object, V_1 and V_2 are visual proto-objects of the simulated participant and the additional person respectively.

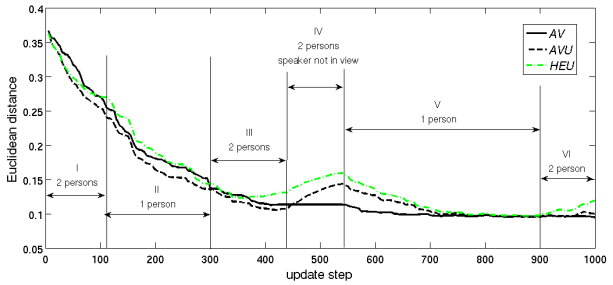

Fig. 7. Scenario 1: Comparison of three methods using Euclidean distance from offline-calibrated maps.

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 $k_n = 19$ are the numbers of positions, cues, frequency channels and nodes respectively.

B. Online scenarios

In online scenarios, audiovisual integration was learned with and without consideration of the situation where the matched visual proto-object is not in the STM respectively, as explained in Section III-C. To simplify the description, Algorithm 1 and Algorithm 2 were denoted as "*AVU*" and "*AV*", respectively. They were also 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* is similar to methods in [17], [18] for linking auditory and visual signals. Audio-motor maps were initialized with random numbers in the range $[-0.5, 0.5]$ using a uniform distribution.

In the first scenario, we simulated a participant with a

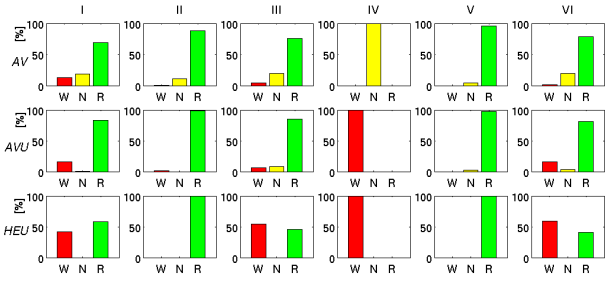

Fig. 8. Scenario 1: Percentage of update steps where a wrong visual proto-object (W), no visual proto-object (N) or a right visual proto-object (R) is chosen for audiovisual integration in each phase.

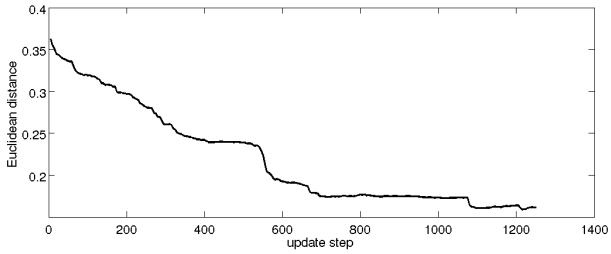

Fig. 9. Scenario 2: Euclidean distance between online-adapted and offline-calibrated maps in a natural dialog scenario. *AV* is used in online adaptation of audio-motor maps.

loudspeaker on which a picture of a face was attached. The loudspeaker was placed on the same position as in offline calibration. An additional person dynamically entered and vacated the room, but did not speak, thus the only sound source is the loudspeaker at 0°. During online adaptation, the robot head oriented itself to a random horizontal angle after an update step was finished. Fig. 6 sketches the experimental setting of the first online scenario. The head angle p_h was in the range $[-90, 90]$ except in phase IV when it was in the range $[-25, 25]$. In phase I, the person was visible in the room ($N = 2$). Then the participant vacated in phase II ($N = 1$) and entered the room again in phase III ($N = 2$). In phase IV the loudspeaker was turned off. The additional person stood outside the camera field of view and talked to the robot ($N = 1$), hence the matched visual proto-object is not in the STM. In phase V the person vacated the room and the loudspeaker was turned on ($N = 1$). Finally, the person entered the room again in phase VI ($N = 2$).

Fig. 7 shows a comparison of the three methods using Euclidean distance from offline-calibrated maps in six phases. Fig. 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 *AV* and *AVU* 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, *AV* and *AVU* 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 *AVU* selected the wrong visual proto-object for audiovisual integration, so that the quality of maps became poor. In comparison, *AV* refused audiovisual integration and the performance of online adaptation did not decrease. We also verified the quality of online-adapted maps using them to localize sounds from the calibration database. The average position error of 300 measurements using online-adapted maps with *AV* 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 second scenario the loudspeaker was not used. Up to four participants talked to the robot alternatingly. They moved

slightly while talking, and dynamically entered and vacated the room. Fig. 9 shows that the Euclidean distance between online-adapted and offline-calibrated maps decreased over time. It is thus demonstrated that we can learn maps in a natural dialog scenario.

V. Conclusion

We have suggested a system for learning audiovisual integration based on temporal and spatial coincidence in scenarios where the current sound is to be related to one out of many visual signals - for instance in the case where a sound is heard and many faces are seen. Firstly, the probabilities that each visual signal is related to the current sound are computed using spatial coincidence. Then we confirm that the visual signal with the maximal probability belongs to the current sound. If one visual signal shows a very high probability and all other visual signals have low probabilities, this indicates a reliable integration. Conversely, when all visual signals have quasi equal probability, the integration is unreliable and is not executed. We have also considered the situation where the current sound is related to a visual signal that has not yet been seen.

The system was tested in online adaptation of audio-motor maps. Since audio-motor maps are not reliable at the beginning of the experiment, learning is bootstrapped using temporal coincidence when there is only one auditory and one visual stimulus. In the course of time the system automatically decides to use both spatial and temporal coincidence depending on the quality of maps and the number of visual sources.

We have shown that our audiovisual integration method performs well when more than one visual source appears. The integration performance does not degrade when the related visual source has not yet been spotted. The audio-motor maps which are online adapted using audiovisual integration can reach the performance of offline-calibrated maps. We have also shown that the online adaptation using our audiovisual integration works in a natural dialog scenario.

Presently, if a visual target disappears for a certain time and then reappears or moves quickly, it will be considered as a new one. This is a shortcoming of using only spatial coincidence to group visual proto-objects in the STM. We plan to employ more grouping features such as color and size in future work. Additionally, if $P_{common}(p_a, p_{v_{N+1}})$ is bigger than the maximal $P_{common}(p_a, p_{v_i})$ ($i \in [1,N]$), the current sound is very probably related to a visual proto-object that is not in the STM. In this case we could trigger a search behavior such as head orientation.

Acknowledgment

This work has been supported by the Honda Research Institute Europe. We thank Andrew Dankers for his support with the final version of the paper.

References

[1] 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.

[2] 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.

[3] E. E. Khoury, G. Jaffre, J. Pinquier, and C. Senac, "Association of audio and video segmentations for automatic person indexing," *International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2007.

[4] M. Radeau, "Auditory-visual interactions in spatial scene analysis: development and neural bases," *Auditory-Visual Speech Processing (AVSP)*, 1998.

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

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

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

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

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

[10] 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.

[11] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analyse*. Wiley-IEEE Press, September 2006.

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

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

[14] ——, "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.

[15] J. Lewald, "Rapid adaptation to auditory-visual spatial disparity," *Learning & memory*, vol. 9, pp. 268–278, 2002.

[16] L. Natale, G. Metta, and G. Sandini, "Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head," *Robotics and Autonomous Systems*, vol. 39, pp. 87–106, 2002.

[17] 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.

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

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

[20] 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.

[21] 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)*, 2007.

[22] 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.

[23] C. Karaoguz, M. Dunn, T. Rodemann, and C. Goerick, "Online adaptation of gaze fixation for a stereo-vergence system with foveated vision," *International Conference on Advanced Robotics (ICAR)*, 2009.

[24] N. J. Butko and J. R. Movellan, "Learning to look," *Proceedings of the 2010 IEEE International Conference on Development and Learning*, pp. 70–75, August 2010.

[25] 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.