# Attention via Synchrony: Making Use of Multimodal Cues in Social Learning

Matthias Rolf, Marc Hanheide, and Katharina J. Rohlfing

***Abstract*—Infants learning about their environment are confronted with many stimuli of different modalities. Therefore, a crucial problem is how to discover which stimuli are related, for instance, in learning words. In making these multimodal "bindings," infants depend on social interaction with a caregiver to guide their attention towards relevant stimuli. The caregiver might, for example, visually highlight an object by shaking it while vocalizing the object's name. These cues are known to help structuring the continuous stream of stimuli. To detect and exploit them, we propose a model of bottom-up attention by multimodal signal-level synchrony. We focus on the guidance of visual attention from audio-visual synchrony informed by recent adult–infant interaction studies. Consequently, we demonstrate that our model is receptive to parental cues during child-directed tutoring. The findings discussed in this paper are consistent with recent results from developmental psychology but for the first time are obtained employing an objective, computational model. The presence of "multimodal motherese" is verified directly on the audio-visual signal. Lastly, we hypothesize how our computational model facilitates tutoring interaction and discuss its application in interactive learning scenarios, enabling social robots to benefit from adult-like tutoring.**

***Index Terms*—Attention, infant-directed communication, multimodality, social learning, synchrony.**

## I. Introduction

IMAGINE you are a stranger in a foreign country. Every time a rabbit runs by, the people say "gavagai." What gives you the basis for the assumption that "gavagai" refers to the rabbit? In pointing to the ontologic relativity, Quine [1] suggested that it can also be the "rabbitness"—as a more abstract or superordinate category—or "rabbit's leg" as a more specific or subordinate category. Imagine you are a child that learns language. You see a round thing that can roll. Your parents say "ball" to it. What gives you the basis for the assumption that "ball" refers to the object and not to the action of rolling? In pointing to this mapping problem, researchers in developmental psychology and linguistics have identified this reference process as central in word-learning theories. A child has to know what the word refers to in order to map, i.e., to remember the appropriate link. Concerning the problematic situation sketched above, Quine teaches us that we cannot resolve the reference through ostension exclusively. His considerations should invite us to look for additional sources for reference resolution. What else if not the ostensive context alone can help to resolve the reference?

Some researchers criticize the view that in the process of learning language, words are just mapped onto existing concepts about objects. Tomasello [2], for example, attacks the false metaphor of mapping and suggests that we lose the exclusive cognitive and associative view on learning in favor of a more social perspective on learning [46]. The social approach is in line with late Wittgenstein's philosophy of language. In this view, it is not about two parts, a word and an object, that need to be linked; it is about the situation, in which a person uses a symbol for the purpose of redirecting another person towards the object (see also [3]). "In fact, if attention is guided, little ambiguity in interpretation need result" [4, p. 720]. In this social approach, it is not only the word as the sole information available to the hearer for the reference resolution. Also the behavior of the speaker and the circumstances of a situation as well as the hearer's experience contribute to the formation of meaning. It is important to note that in the social-pragmatic approach to learning, there is still a place for word-referent correlations. The critique targets more the exclusivity of the mapping process. It suggests instead that the attentional processes of the participants involved in the situation are important as well. Furthermore, we are convinced that social learning is in its fundamentals driven by perceptual cues that convey learning-relevant information. This is why we take a closer look at the signal level in social learning scenarios.

### A. Multimodal Motherese

Recent approaches to word learning take advantage of interactive processes between participants. Zukow-Goldring [5], [46] investigated naturalistic situations, in which mothers were interacting with their children. She suggested that the participants develop a sense of shared understanding of actions, on which basis children learn language. Gogate *et al.* [6] took an experimental approach and investigated how mothers interact with their children when they try to teach them a new word. In this study, two new words were given to two new objects and two words referred to actions with the new objects. In coding the mothers' action, the authors discriminated between situations in which target words were presented i) in synchrony with a motion of the referent object, ii) asynchronously to an object motion, iii) without any object motion, and iv) while the infant was

Manuscript received December 13, 2008; revised March 30, 2009. First published April 17, 2009; current version published May 29, 2009. The work of K. J. Rohlfing was supported by the Dilthey Fellowship (Volkswagen Foundation) and by the European Community under the Innovation and Communication Technologies program of the 7th Framework for ITALK project under Grant ICT-214668.

M. Rolf is with the Research Institute for Cognition and Robotics, Bielefeld University, Germany (e-mail: rolf@CoR-Lab.Uni-Bielefeld.de).

M. Hanheide and K. J. Rohlfing are with the Research Institute for Cognition and Robotics and the Applied Informatics Group, Bielefeld University, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2009.2021091

holding the referent object. For the goal of the paper, it is important to emphasize that in this semi-experiment, a manual coding system was used for the analysis. The authors found out that when mothers were asked to teach a new word for the objects or actions, they moved the objects in temporal synchrony with the new label. They observed that when the children were more experienced with language (starting from 21 months of age), the synchronous action was less pronounced in their mothers. Gogate *et al.* [6] argue that preverbal and early verbal children need temporal synchrony in order to make the reference to the object or its action on the one hand. On the other hand, these data also show that parental behavior responds to the needs of the child. Thus, the multimodal communication towards infants is adapted to infants' increasing abilities to find out word–object relations on their own.

A further experiment by Gogate and Bahrick [7] investigated seven-month-olds' abilities to map a syllable onto an object. Infants were presented the new object moving either in temporal synchrony with the new label or in an asynchronous way. Infants could remember the label of the objects after 10 min only when they saw the stimulus in a synchronous condition. The authors interpret the results as consistent with the view that prior to symbolic development, infants learn and remember word-object relations by perceiving redundant information in the vocal and gestural communication of adults. Against this experimental background, Gogate *et al.* [8] propose a basis for the understanding of spoken words: It is the "early detection of intersensory relations between conventionally paired auditory speech patterns (words) and visible objects or actions." As Tomasello [2] suggested, Gogate *et al.* view the learning process as dynamic and in reciprocal interaction, supported strongly by general intersensory perception and selective attention.

### B. Intersensory Redundancy

The *intersensory redundancy hypothesis* (IRH) [9] attempts to explain how synchrony of signals can guide infants' selective attention and contribute to the learning process. The starting point is the fact that humans perceive the environment over various modalities like vision, touch, and hearing. For example, when driving a car, one sees the street and other cars, hears motors or maybe sirens, and feels the wheel and the gearshift. Consequently, we can represent our environment in terms of several modalities at the same time.

Since the 1990s, it is increasingly realized that not just the isolated modalities but also their interplay drives our perception and cognition. Multimodality provides a unary, integrated understanding of our environment [9]. According to recent research, integration across modalities occurs before each stimulus is fully processed unimodally [10], [11]. In this interplay, stimuli from one modality can help interpreting stimuli from another modality that are ambiguous on their own. A well-studied example is lip-reading. Here, the view of the mouth can enhance the recognition performance of the heard words [12], [13]. What binds the modalities together and helps to interpret the incoming signal are amodal properties such as synchrony, rhythm, or intensity [9]. For example, a crashing glass produces a sharp stimulus in vision and hearing, appearing at the same time. Thus, they are temporally synchronous. Synchrony has been demonstrated to affect attention: Signals that reinforce each other on the basis of amodal properties promote earlier processing. They thus attract the attention of perceivers and become foreground in contrast to other properties to become background [9]. The power of cross-modal binding has been shown for newborn and young infants as auditory stimulation has been found to facilitate the visual attention [14], [15]. Infants as young as two months are sensitive to voice–lip synchrony during speech [16]. Furthermore, recent studies by Zukow-Goldring *et al.* [17] using eye-tracking technology with video confirms that infants 9 to 15 months old prefer looking at objects that are presented in a synchronous word–object condition. In addition, children showed a better comprehension of the word when it was uttered matched with the rhythm of the object movement. The infant's initial sensitivity to amodal information such as synchrony—as has been shown in the study by Gogate *et al.* [6]—provides an economical way of guiding perceptual processing to focus on meaningful, unitary events [9].

Even though it was experimentally shown that parents modify their behavior and show an object in synchrony with its label [6], little is known about whether adults synchronize their behavior in other than word-learning contexts. Yet the problem of reference is considered to be similar. It is plausible to assume that while synchronizing actions from different modalities, parents increase the saliency of specific action segments [18] and provide structure to the input, from which children learn. It seems like the sensitivity towards amodal properties on the learners' side and the modified behavior that provides lots of amodal overlap between modalities on the tutor side, is a crucial part in social learning scenarios. They can be seen as one key to the reference resolution problem sketched above.

### C. Approach

In designing artificial systems, we have to consider such symbiotic way of interaction and attempt to systematically take advantage of them [19]. Within the set of mechanisms that will be needed to enable robots to learn from humans in a socially interactive way, we take the perspective of the infant and *model its sensitivity to synchrony* as an attention guiding cue provided by the tutor. Thus, we are constructing attention via synchrony. As sketched above, this guidance has been identified as a crucial ingredient of social learning. However, before synchrony cues can be used in *any* way, they have to be detected by the robot's perceptual system. In our approach, we focus on the *detection* of synchrony between auditory and visual information. In our notion, detection is about discriminating synchrony and asynchrony in the spatiotemporal domain in the course of interaction. In order to make use of synchrony for the guidance of attention, we use low- or rather signal-level information. Those features are known to be available at preattentive stages [20] of cortical processing and are plausible to be available even on early stages of development. We interpret the detected synchrony directly in terms of attention: visual stimuli that are most synchronous to the audio domain receive the most attention. We describe our computational model for synchrony detection in Section II. In Section III, we present an experiment, where we tested the receptiveness of our model to child-directed cues in parents' tutoring and its applicability for the guidance of attention. Lastly, we discuss our method and results in the context of interactive social learning and robotics in Section IV.

## II. Computational Model

Research on temporal intermodal perception of synchrony mostly builds on the notion of events [6], [21], [22]. While most studies lack a precise definition of "synchrony," a common formalization of the term is completely missing [23]. In particular, the temporal relations between two modalities become complicated when not just a single event on each modality has to be considered. Lewkowicz *et al.* [21] describe a whole hierarchy of relations across simultaneity/synchrony, shared intensity, durations, and rhythms. All of those concepts are not trivial to formalize, particularly when not each event has a counterpart in another modality. For instance, Matatyaho *et al.* describe forward arm movements, simultaneous with speech labels for a held object, while backward movements in between have no counterpart in the speech domain [24].

Here, we go in line with [22] and [25] and basically regard synchrony as the property of two events to occur at the same time—i.e., they have simultaneous on- and offsets. However, this does not immediately yield a good formalization in the case of multiple events in each modality and not exactly simultaneous timing.

The notion of an event is likewise hardly formalized. While Gogate *et al.* [6] define events mainly according to discrete word on- and offsets, this notion is hardly applicable when facing a nonsegmented, continuous perceptual stream. But basically, we must consider any visual and auditory perception as a nonstructured stream and define synchrony in order to end up with a structured representation following a "developmental pathway." The challenge is thus to define synchrony measures on low-level and preattentive features. There is, for instance, broad evidence that features like color or simple motion are computed massively parallel all over the visual field [20] before attention is actually constructed [26].

At that level, *temporal correlation* between signal flows can be seen as direct adoption of synchrony to continuous signals: stimuli gain high correlation when the signal values decrease or increase simultaneously in several modalities. In fact, stimuli can also gain gradual correlation when a small temporal delay is introduced between them, as long as the delay is smaller than the basic period length of the stimuli. At first, this notion is conceptually orthogonal to the notion of synchrony defined for events: i) events in two modalities do not necessarily gain signal correlation due to the concrete shape of the continuous signal and ii) also in the absence of anything considered as event there can be correlation. However, the differences depend on both the definition of an event and the choice of features. Events can, e.g., be defined as peaks in signal values [27]. Also, one can use features that directly indicate events. Thus temporal correlation provides a generic formal notion of synchrony that is natural at signal level but also applicable at event level.

### A. Synchrony Detection

Our method is based on an algorithm proposed by Hershey and Movellan [23]. The algorithm detects temporal correlations (synchrony) between visual features and auditory features. Therefore, each image location (i.e., pixel) is treated separately. The statistical analysis is restricted to a small window in time that is shifted over the audio/video stream. Since each pixel yields independent estimates of synchrony, the result is a topographic map of synchrony. As the final synchrony estimate is a mutual information measure, such maps are also referred to as "mixelgram" (see Fig. 1). Like many other approaches to signal-level synchrony, the algorithm was originally developed for statistical sound-source localization (see [28] for an overview). In that scenario, it is assumed that physical sound-sources provide synchronous patterns across modalities. Stimuli that provide synchrony but do not correspond to an immediate sound-source are considered as false positives or disturbances. However, our application context is broader since we want to detect social cues that do not directly refer to physical sound sources. For our purposes, the algorithm has two important properties.

1) The algorithm contains no assumptions about the kind of visual or auditory stimuli. It is for instance not specific to human faces and voices. From the learning perspective, this is important since such specific patterns shall be the result of, but not a prerequisite for, an overall learning process.
2) The algorithm is fast enough to detect synchrony in real time with reasonable video resolutions and sampling rates. The goal is to employ the method also in artificial intelligent systems like robots as discussed in Section IV. Here, the method is to be used within a closed interaction loop.

The model has already been compared to infants' abilities in synchrony detection by Prince *et al.* [29]. They showed that the method can indeed model the infants abilities in some situations. However, we make use of more sophisticated methods for filtering and quantitative synchrony estimation, which we describe in the following.

*1) Statistical Estimation:* The basic mathematical assumption for the statistical analysis is that the values of visual and auditory features originate from a joint probabilistic process. This process is assumed to be stationary and Gaussian for a short period of time

$$\left\{ \begin{pmatrix} a(t) \\ v(x,y,t) \end{pmatrix} \right\}_t \sim \mathcal{N}(\mu(x,y), \Sigma(x,y)).$$

Here we denote the set of $n$ audio features over time as $a(t) \in \mathbb{R}^n$ and the set of $m$ video features for each pixel $v(x,y,t) \in \mathbb{R}^m$. $\mathcal{N}$ is the joint ($n+m$-dimensional) Gaussian distribution with mean $\mu$ and variance $\Sigma$.

For the synchrony detection, the parameters $\mu$ and $\Sigma$ are estimated from the video data $\{a(t_k), v(x,y,t_k)\}_k$, where $k \in \{1,\ldots,T\}$ denotes the index of a video frame and $T$ the number of frames in a video. Hershey and Movellan suggested to estimate $\mu(x,y,t_k)$ and $\Sigma(x,y,t_k)$ over a time window of $s$ frames $\{a(t_l), v(x,y,t_l)\}_l$, with $l \in \{k-s+1,\ldots,k\}$. For practical reasons, we do not use a hard time window but compute exponentially smoothed estimates of $\mu(x,y,t_k)$ and $\Sigma(x,y,t_k)$. The data from a current frame $(a(t_k), v(x,y,t_k))$ receive a constant

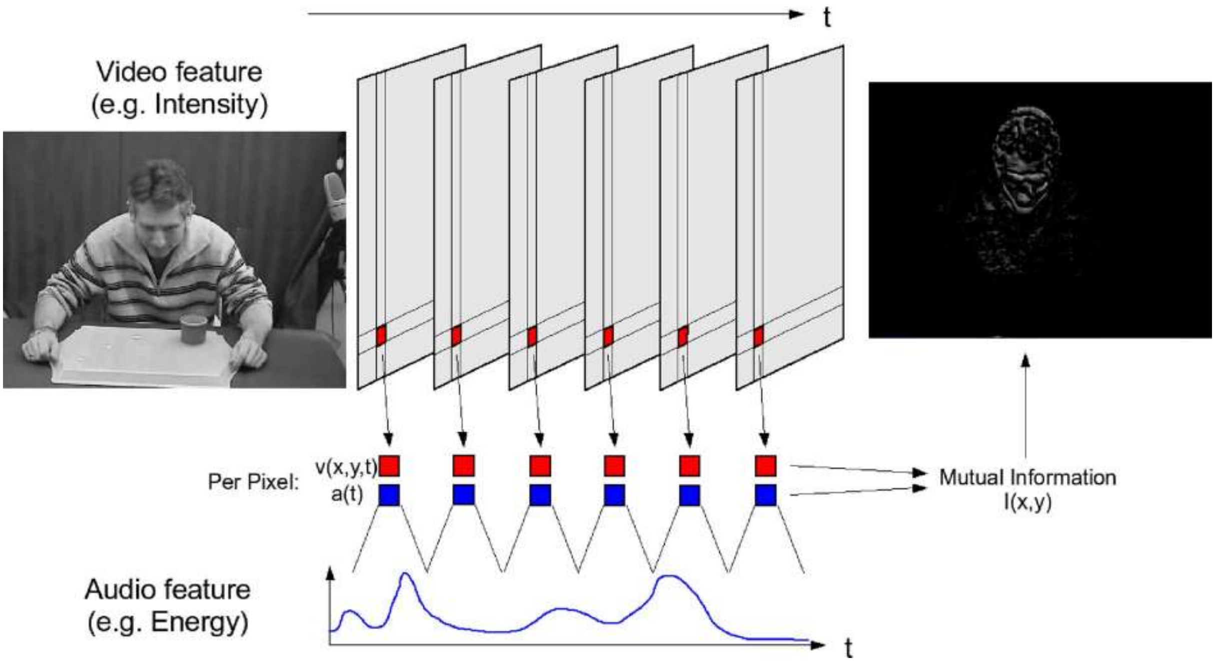


Fig. 1. Schematic overview of synchrony detection: a video feature is per pixel compared to a audio feature for a short window in time. Mutual information is computed based on a linear correlation coefficient, yielding a topographic map of synchrony.

weight $\alpha \in ]0;1[$ and are recursively combined with the previous estimates of mean and variance

$$\mu(x,y,t_k) = \alpha \cdot \begin{pmatrix} a(t_k) \\ v(x,y,t_k) \end{pmatrix} + (1-\alpha)\cdot \mu(x,y,t_{k-1})$$

$$\Sigma(x,y,t_k) = \frac{1}{1+\alpha}\left(\alpha \cdot \left(\begin{pmatrix} a(t_k) \\ v(x,y,t_k) \end{pmatrix} - \mu(x,y,t_{k-1})\right)^2 + \Sigma(x,y,t_{k-1})\right).$$

This update rule has several advantages: first, this scheme allows very efficient frame-to-frame computation without bookkeeping of other frames than the current one. Secondly, it yields smoother estimate characteristics over time: though rapidly changing new stimuli cause rapid changes of the estimates, forgetting is gradual and does not cause discontinuities. Thirdly, it is also smooth in the influence factor $\alpha$, which is not the case for varying values of $s$ for a fixed time-window. Small changes in $\alpha$ only reweight the past samples slightly, without making abrupt changes.

The estimates of (co)variances $\Sigma(x,y,t_k)$ are then used to express the degree of synchrony between audio and video in terms of mutual information $I$. Assuming a Gaussian distribution yields an immediate relation

$$I_{A,V}(x,y,t_k) = -\frac{1}{2}\log\left(\frac{|\Sigma_A(t_k)|\cdot|\Sigma_V(x,y,t_k)|}{|\Sigma(x,y,t_k)|}\right)$$

$$\Sigma(x,y,t_k) = \begin{pmatrix} \Sigma_A(t_k) & \Sigma_{A,V}(x,y,t_k) \\ \Sigma_{A,V}(x,y,t_k)^T & \Sigma_V(x,y,t_k) \end{pmatrix}.$$

In the case of each audio and video feature ($n = m = 1$), this relation can be simplified [23] and expressed in terms of a Pearson correlation coefficient $\rho$

$$I_{A,V}(x,y,t_k) = -\frac{1}{2}\log\left(1-\rho^2(x,y,t_k)\right)$$

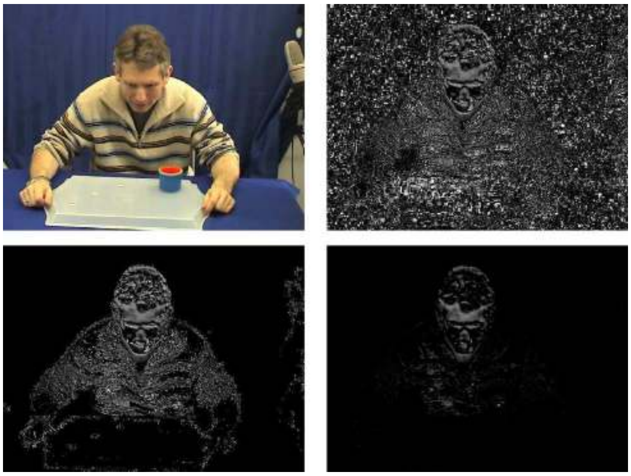


Fig. 2. (Top left) Original RGB-frame from a test video. (Top right) Mutual information obtained with image intensity and audio energy as features and $\alpha = 0.05$ at 25 fps. Background noise causes intensive correlation artifacts. In the illustration, white corresponds to a mutual information of 0.51, which is a Pearson correlation of ±0.8. Black indicates zero correlation. (Bottom left) A threshold of $\mathcal{T}_V = 10$ on video variance. Static background pixels are mostly excluded. (Bottom right) An additional morphological erosion removes remaining background pixels and also outstanding noise pixels in regions with activity.

$$\rho(x,y,t_k) = \frac{\sigma_{A,V}(x,y,t_k)}{\sqrt{\sigma_A(t_k)\cdot\sigma_V(x,y,t_k)}}.$$

Thereby $\sigma_{A,V}$, $\sigma_A$, and $\sigma_V$ are the now scalar estimates of variances and the covariance of audio and video feature.

The overall result is one mutual information image (mixelgram) per frame. High values of mutual information are visualized with lighter grayscale values (see Fig. 2) and express a high degree of synchrony between audio and video. In the original scenario of sound-source localization, high mutual information reflects a possible sound-source at a certain image location.

However, in our scenario, we are less interested in such physically causal correlation. Instead, we rather try to investigate the role of synchrony for attention in tutoring. Hence, it is assumed that the model is also perceptive to synchrony induced by the tutoring process itself. Therefore, a mixelgram can directly be interpreted in terms of attention so that image regions with high mutual information receive the highest degree of attention.

*2) Filtering:* Pearson's correlation and mutual information as measures of interdependence between audio and video indicate the significance of a relation between both modalities. It is noteworthy that the significance of the signal is not taken into account, since they are independent of shift and scale of feature values. In fact, most pixels in an image are usually static apart from noise, thus providing no significant change over time. Nevertheless, those pixels can cause high correlation just by chance (see Fig. 2). This behavior is not desirable since noise should not be the driving force in attention.

We propose a two-stage filter process to exclude insignificant visual stimuli and noise. The first stage excludes pixels without activity. As measurement of activity, we use the variance over time on each pixel. Note that the variance is already available due to the correlation estimation. If the variance on a pixel is below a specified threshold $\mathcal{T}_V$, mutual information is set to zero. Fig. 2 illustrates the effect: large areas of stationary background are filtered out. Still, there is notable noise in regions that must be considered to be active. This noise results in single, outstanding pixels with high mutual information [Fig. 2 (bottom left)]. These single pixel distortions are effectively handled by the second filter stage: a morphological erosion. Each pixel value is replaced by the minimum value of its direct neighborhood. Thereby, single outstanding pixels are completely erased, while massive regions of mutual information are retained [Fig. 2 (bottom right)].

## B. Quantitative Analysis

An empirical problem in using this method is that it does not yield an immediate estimate of overall synchrony contained in a video. This problem was already addressed in [29]. However, the proposed solutions appeared to be very specific to the kind of incoming stimuli to overcome the noisy properties of the mixelgrams. Due to our filtering procedure, we can apply a straightforward averaging method to get quantitative estimates of synchrony.

*1) Synchrony Measurement:* As a first step, each mixelgram is condensed to a scalar estimate of synchrony $S(t_k)$ for that time in the video. Therefore, we average over the set of pixels $\mathcal{P}(t_k)$ that was not set to zero within the filtering process

$$S(t_k) = \frac{1}{|\mathcal{P}(t_k)|} \sum_{(x,y)\in\mathcal{P}(t_k)} I(x,y,t_k) \tag{1}$$

$$\mathcal{P}(t_k) = \{(x,y) : 0 < I(x,y,t_k)\}. \tag{2}$$

The only assumption that has to be made is that a sufficiently large number of pixels is not filtered out—which is the case when motion is present in the visual appearance of the scene. This assumption is consistently fulfilled in our experimental study. Finally, the values of $S(t_k)$ are averaged over time to gain an estimate of synchrony for a whole video

$$\mathcal{S} = \frac{1}{T}\sum_{k=1}^{T} S(t_k). \tag{3}$$

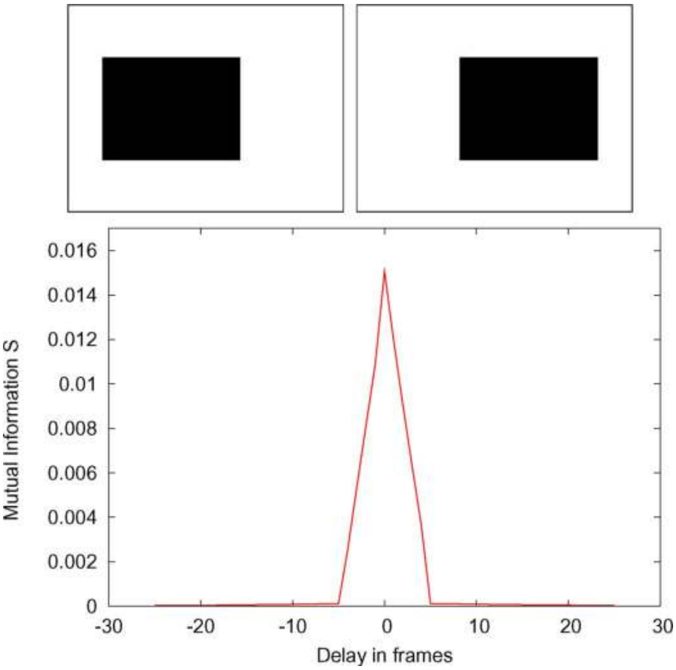


Fig. 3. The artificial test video contains a black rectangle that moves from left to right after staying on the left side for 2 s. The movement takes five frames and goes along with a beep. The combination of different visual features and audio energy provides the expected synchrony maximum at a zero delay.

*2) Example:* As a basic test case for the synchrony estimation, we used an artificial test video containing a black rectangle that makes a horizontal movement across five video frames (see Fig. 3). The audio channel contains a beep tone that is exactly aligned in time with the rectangle movement. Thus audio and video are highly synchronous. In order to test the synchrony estimation, we adopted an experimental scheme from [30]: the audio stream is delayed against the video stream, yielding a gradual desynchronization of both modalities. The estimated degree of synchrony is then compared for different delays, whereas the maximum of synchrony should be located at delay 0. Fig. 3 shows that the measure $\mathcal{S}$ [(3)] is basically able to discriminate between synchronous and asynchronous conditions. Here we used difference images as video feature, audio energy for the sound modality, and a smoothing factor $\alpha = 0.05$.

*3) Normalization:* The proposed measure $\mathcal{S}$ is basically an average value of mutual information, where the size of active image regions $\mathcal{P}(t_k)$ is compensated. The measure yields a good basis to compare highly related stimuli with respect to their synchrony (as shown in Fig. 3) or to compare different parameters or features. However, our empirical goal is to compare different videos or rather scenes. Though they are comparable on a semantic level, they show radically different patterns at signal level. Therefore, we introduce an additional normalization step: in a separate computation, the original audio track is replaced by

Gaussian white noise. Then synchrony is measured in the same way as with the original audio track. The result $\mathcal{S}^{\text{noise}}$ provides a baseline for the synchrony with the original audio track. Since $\mathcal{S}^{\text{noise}}$ is per definition a stochastic measure, the average across $n$ instantiations of Gaussian noise is used (here $n=10$). Then we use the ratio between both estimates as measure of synchrony contained in a video

$$\mathcal{S}^{\text{relative}} = \frac{\mathcal{S}}{\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}^{\text{noise}}(i)}. \tag{4}$$

This measure indicates the mutual information gain of the original audio track, relative to pure audio noise—and thus how synchronous audio and video are. Note that $\mathcal{S}^{\text{noise}}$ is neither zero nor constant across different videos—also not its expectation over different instantiations of noise. The independent Gaussian distribution of the white noise provides indeed no mutual information to any other probability distribution. However, the algorithm deals with sampled values of both the audio and the video signal; thus correlation can occur by chance. Also, the video signals do in practice not fulfill the two assumptions made in correlation computation: that the samples are i) independently chosen ii) from a Gauss-distribution. Therefore, $\mathcal{S}^{\text{noise}}$ measures how prone the visual patterns are to spurious correlations, which can be caused by violations of those assumptions. Thus it provides a clear baseline for the mutual information, and normalizing against it yields a more interpretable and robust measure.

## III. Experiment

The major goal of this experiment is to investigate synchrony in a social learning scenario in terms of child-directed communication. It is important to note that the data for our analysis encompass a contingent interaction since we analyzed parental behavior during a real situation with their children. In this situation, they continuously reacted and adapted to their child. The basic hypothesis is that during a demonstration, parents provide additional learning cues by synchrony. The hypothesis is tested by comparing the degree of audiovisual synchrony between adult- and child-directed communication. In order to use such synchrony cues provided by a tutor, it is important to understand what is synchronous and when. Therefore, we discuss several examples of the spatial distribution of synchrony.

### A. Scenario and Setup

*1) Participants:* We investigated 184 videos showing 48 participants. The data stem from the original video corpus (also used in [15] and [16]), which contains videos of 66 parental couples interacting with their children. The infants' age ranged from 8 to 30 months. For this analysis, the selection of the subjects was restricted to setups that comprise both mother and father, demonstrating the interaction to both partner and child (four runs) without disturbance. The excluded videos included the experimenter walking through the scene, the infant pulling the tablecloth down, verbal interaction with the experimenter, and crying from the infant. Further, only videos with an existing speech annotation were used.

After all the selection process, 192 videos were selected for the analysis. They were equally distributed over four tasks, each with 12 parental couples in four runs. For the salt shaker demonstrations, only 11 parental couples could be used due to missing speech annotations. Further, two videos (one bell-task and one salt shaker, both child-directed) were excluded due to a corrupted audio track. As a direct comparison was impossible for these two videos, the corresponding adult-directed demonstrations were also excluded, yielding a final number of 184 videos available for analysis. The videos thereby show 24 different parental couples and thus 48 different subjects. All videos were analyzed in the original resolution of 720 × 576 pixels at 25 fps applying our proposed computational model. The audio tracks contain mono sound, sampled with 44 100 Hz.

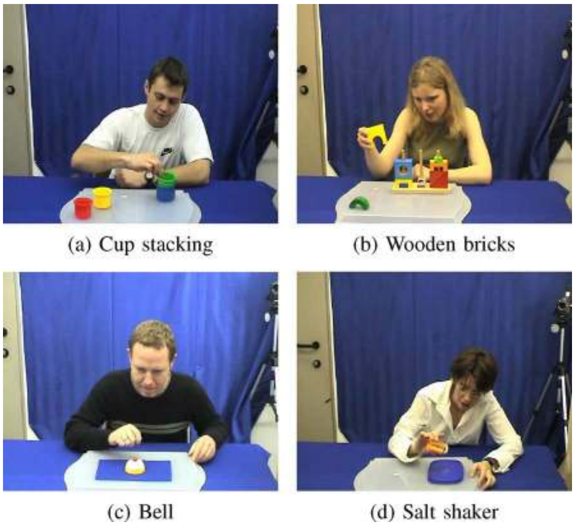

(a) Cup stacking
(b) Wooden bricks
(c) Bell
(d) Salt shaker

Fig. 4. Investigated scenarios of multimodal motherese: parents demonstrate different object interactions to either their child or their partner.

*2) Materials:* For this analysis, four tasks were selected (see Fig. 4): the cup stacking, in which parents demonstrated how the cups can be stacked; the wooden bricks, in which tasks parents were instructed to put a block on a pole (altogether three blocks were put); the bell, which rang after pressing the red button; and the salt shaker, which was filled with salt, with the parents demonstrating how to shake the salt on a the blue tray.

*3) Procedure:* In the study, both parents interacted with their child and with an adult. The first run was an adult–child interaction, in which one parent (randomly selected) and her or his child sat across the table. The parent was instructed to demonstrate the function of the objects to the child. Here, the parent was free to teach the word, the action, or both (those two acts were in fact mostly inseparable in the collected data). We asked to move the white tray and to give the objects to the child only after the demonstration. The child was attending to the demonstration and interacting with the parent. In a following adult–adult interaction, the same parent was asked to demonstrate the object to her or his partner. In the third run, the second parent demonstrated the objects to the child. In the fourth run, the same parent demonstrated the objects to an experimenter.

*4) Measurement and Features:* As basic measure of synchrony for our experiment, we used $\mathcal{S}^{\text{relative}}$ [see (4)]. The videos selected for analysis still contain a wide range of visual and acoustic disturbances. Therefore, we restricted the averaging of mutual information over time to those frames that

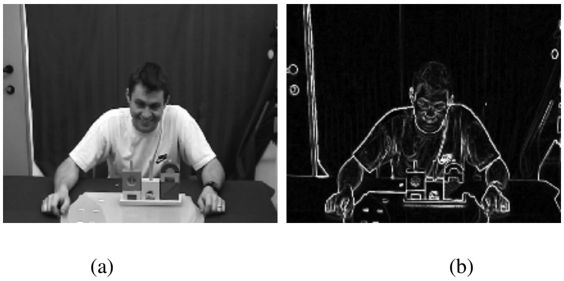


Fig. 5. (a) Grayscale image and (b) gradient strength image conducted from Sobel edge filters.

show the parent speaking. For our empirical purposes, this is a reasonable method, since we are interested in the parents behavior with respect to synchrony between speech and motion signals.

As audio feature, we consistently used audio energy (as used in [23] and [31], or similarly root mean square values in [29]), which is the average squared sample value within a chunk of audio data. As video features, we used image intensity (grayscale values) and gradient-strength images alternatively (see Fig. 5). In a preliminary study, we found these two image features to gain the best discrimination performance between synchronous and asynchronous stimuli. In contrast, we could not find reasonable discrimination for dynamic features like difference images or optical flow, which are traditionally argued to be better suited to audio data [28], [32]. We tested both intensity and gradient-strength images with three temporal smoothing factors $\alpha \in \{0.02, 0.05, 0.1\}$. The variance threshold was defensively chosen and fixed at $\mathcal{T}_V = 5.0$ for both features.

### *B. Results*

*1) Adult- Versus Child-Directed Tutoring:* The goal of this experiment is to compare synchrony in adult- and child-directed communication. For this purpose, each video showing an adult–adult (AA) interaction is compared to the corresponding adult–child (AC) video. Fig. 6 exemplarily shows the synchrony results for gradient-strength feature and $\alpha = 0.05$. Each point in the plots corresponds to a pair of an AA and AC video, where the synchrony in the AA video is plotted on the $x$-axis and the synchrony in the AC video on the $y$-axis. The first observation is that all except for three videos gained synchrony values above 1.0. That means that the video signals gained higher mutual information with the original audio track than with audio noise. Hence a real synchrony could be detected.

For a direct comparison between AA and AC conditions, the main diagonal (i.e., $x = y$) is shown in the plot. A point above this diagonal indicates that more synchrony is found in the child-directed interaction than in the corresponding adult-directed situation. Indeed, most points (here 62 out of 92) lie above the diagonal. Both median and mean show higher synchrony for the child-directed situation. For this parameter setting, the median synchrony is 2.32 for AA videos and 2.68 for AC videos. The significance of this effect was tested with a two-tailed sign test. The sign test between paired random variables $(a_i, b_i)_{i=1,\ldots,N}$ thereby tests the null hypothesis $H_0 : P(A < B) = P(A > B) = 0.5$. Here the null hypothesis is that synchrony in AC has the same probability to be higher or lower than in the corresponding AA situation. On the dataset presented here, this null hypothesis can be rejected with high significance (error probability $p < 0.001$). The effect can be reproduced across diverse parameter settings (see Table I). In all tested settings, the median of AC synchrony exceeds the median of AA synchrony. The effect also reaches significance in most of the settings. For longer time windows ($\alpha = 0.02$) the effect starts to vanish, and also the values of $\mathcal{S}^{\text{relative}}$ decrease. This indicates that generally less significant synchrony is found, which can be caused by a mixing of time-frames with positive and negative correlation [23], resulting in a close-to-zero linear correlation.

With respect to the different interaction tasks (bottom of Fig. 6), the wooden brick scenario shows a significant ($p < 0.01$) trend towards more synchrony in child-directed communication. For all scenarios, the median of AC synchrony is higher than the AA median, indicating that the effect is rather task-independent.

*2) Correlation AA/AC:* An additional observation in the results is that synchrony in child-directed situations is positively correlated with the synchrony in corresponding adult-directed situations. Due to individual differences, parents tend to produce high synchrony in AA situations (relative to other participants AA synchrony) when they also produce relatively high synchrony in AC situation. We measure this effect with the Spearman rank correlation coefficient. Analogous to Pearson's correlation, it indicates positive correlation with values between 0.0 and 1.0 but is more robust to outliers. For the settings shown in Fig. 6, Spearman's correlation is 0.480. The effect shows to be significant with respect to the null hypothesis that the variables are uncorrelated ($p < 0.01$ with a two-tailed t-test). Also this effect can be reproduced across several parameter settings (see Table II). Thereby a positive correlation is also found within each task. In the setup shown in Fig. 6, rank correlation ranges from 0.34 ($p < 0.1$) in the bell task to 0.51 ($p < 0.02$) in the cup-stacking task. So the overall correlation is not an artifact of the different task means.

*3) Spatial Distribution:* If multimodal motherese provides additional learning cues due to synchrony, it is important to understand *what* these cues actually indicate. In this section, we discuss some exemplary scenes with respect to the spatial distribution on mutual information in the video sequences. Child-directed tutoring was already investigated [18] with respect to the spatial distribution of visual saliency [33]. Thereby, a part of the same cup-stacking demonstrations towards infants was investigated as used in this work. The most salient image position in each frame was categorized as parents' face, parents' hands, demonstrated object, and any other image location. It was shown that different motion patterns in adult- and child-directed communication caused higher saliency on demonstrated objects in child-directed situations. These results indicated which features and parts of the scenes might be relevant. However, the respective location of the maximum is often not congruent with the focus of the current action loci within the tutoring process. In order to analyze whether the proposed model of mutual information can provide a richer attentional cue in this scenario, it is therefore important to understand *where* mutual information is located and how it differs from purely visual cues.

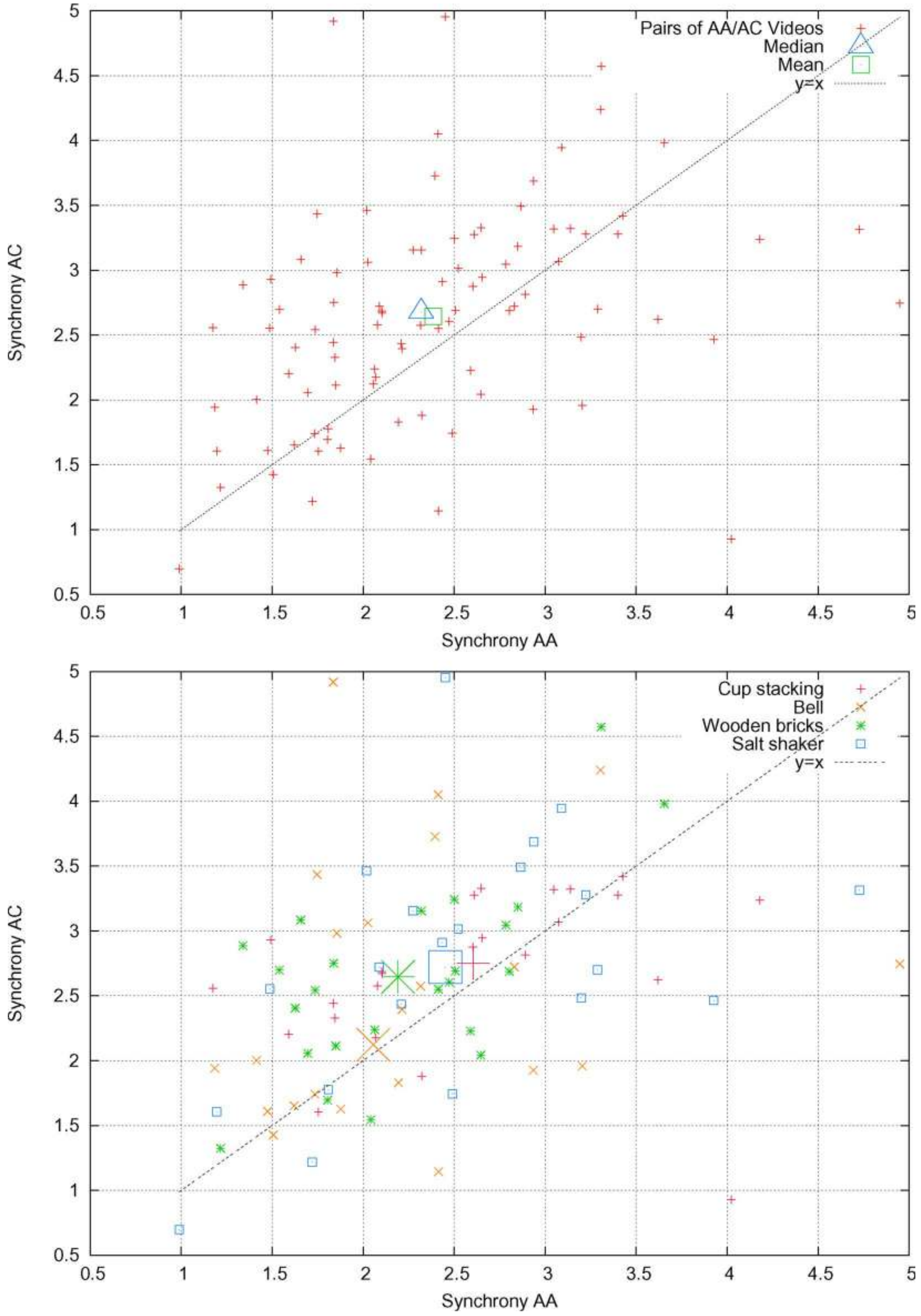


Fig. 6. The plots show the synchrony results for gradient-strength as feature, $\alpha = 0.05$, and $\mathcal{S}^{\mathrm{relative}}$ as measure. Synchrony in each adult–adult video is plotted against the synchrony in the corresponding adult–child video. Both plots show the same data set, where the second plot shows the different object interactions. The median is shown as larger point for each category.

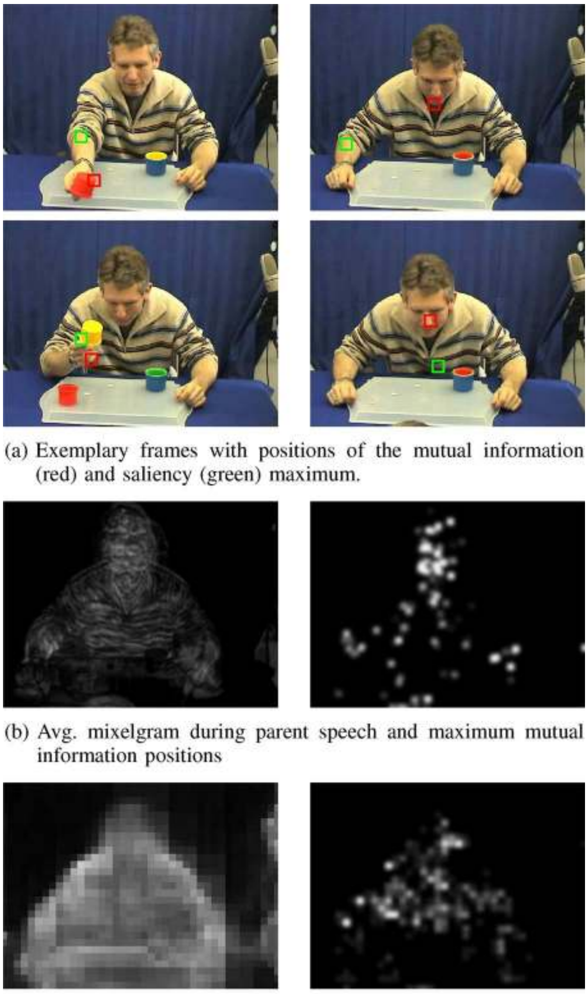
A comparison between attention via saliency and synchrony can generally be done in two ways: first of all, the entire saliency map (or the mixelgram) can be interpreted in terms of *covert* attention [34]. As the potential importance of each image region is encoded in those maps, one can directly compare, e.g., face and hand of a subject with respect to their importance relative to each other. A more condensed view can be gained in terms of *overt* attention [34]: each saliency map and each mixelgram is reduced to a single attended position—a focus of attention. For saliency maps, this is simply the position with the highest value. Thereby we basically used the same saliency configuration as in [18], evaluating intensity, color, orientation, difference images,

TABLE I
COMPARISON BETWEEN ADULT-DIRECTED AND CHILD-DIRECTED COMMUNICATION FOR INTENSITY IMAGES AND SOBEL-BASED GRADIENT STRENGTH AS FEATURES AND DIFFERENT VALUES OF $\alpha$. THE MEDIAN FOR AC-SYNCHRONY IS HIGHER THAN FOR AA IN ALL SETTINGS. THE LAST COLUMN SHOWS THE SIGNIFICANCE WITH RESPECT TO THE NULL HYPOTHESIS $H_0 : P(\text{Sync}_{\text{AA}} < \text{Sync}_{\text{AC}}) = P(\text{Sync}_{\text{AC}} < \text{Sync}_{\text{AA}}) = 0.5$

| | | Median | | |
|---|---|---|---|---|
| Settings | | AC | AA | Significance level |
| int | 0.1 | 3.48 | 2.96 | 0.005 |
| int | 0.05 | 3.86 | 3.09 | 0.001 |
| int | 0.02 | 2.73 | 2.57 | – |
| grad | 0.1 | 2.31 | 2.00 | 0.001 |
| grad | 0.05 | 2.68 | 2.32 | 0.001 |
| grad | 0.02 | 2.18 | 1.96 | 0.1 |

TABLE II
SPEARMAN RANK CORRELATION COEFFICIENTS FOR ALL TESTED PARAMETER SETTINGS. ALL SETTINGS SHOW A SIGNIFICANT POSITIVE CORRELATION BETWEEN SYNCHRONY IN ADULT- AND CHILD-DIRECTED SITUATIONS

| Settings | | Spearman's $\rho$ | Significance level |
|---|---|---|---|
| int | 0.1 | 0.404 | 0.01 |
| int | 0.05 | 0.418 | 0.01 |
| int | 0.02 | 0.399 | 0.01 |
| grad | 0.1 | 0.378 | 0.01 |
| grad | 0.05 | 0.480 | 0.01 |
| grad | 0.02 | 0.227 | 0.05 |



(a) Exemplary frames with positions of the mutual information (red) and saliency (green) maximum.

(b) Avg. mixelgram during parent speech and maximum mutual information positions

(c) Avg. saliency during parent speech and maximum saliency positions

Fig. 7. Subject demonstrating cup stacking towards an infant.

and optical flow by means of Itti and Koch's saliency map model [33].

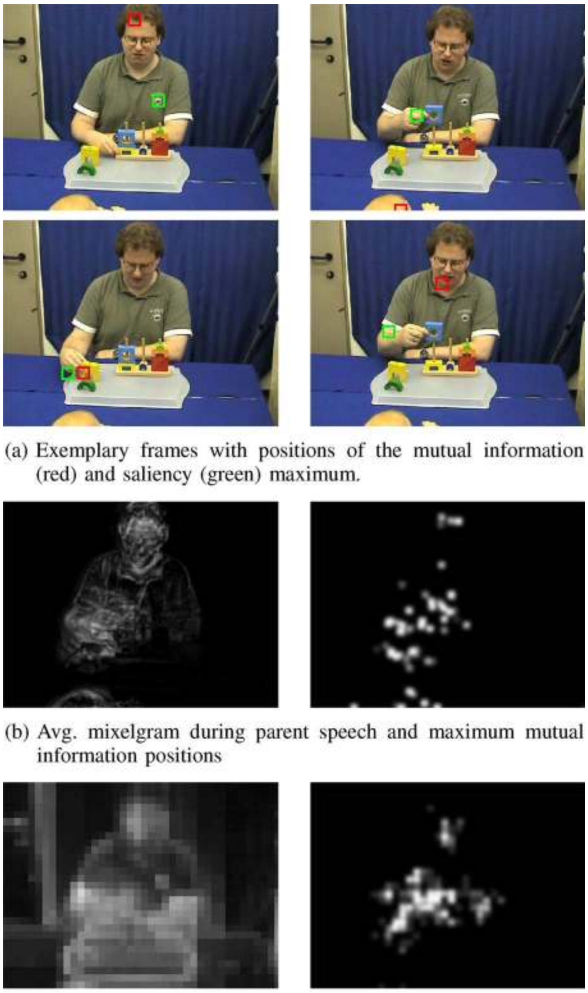
In contrast to [23] and [29], we do not find this location within a mixelgram by means of a center of gravity since we are not interested in a huge region of synchrony but in a region of high synchrony—whatever size it has. However, a pure maximum-pixel detection is not reasonable, since such a pixel does not necessarily reflect a robust maximum in the image region. Here we apply a $15 \times 15$ Gaussian filter to yield a smooth spatial behavior before detecting the maximum. The pixel with the highest value can thus be assumed to reflect a robust maximum.

The analysis of two exemplary videos is shown in Figs. 7 and 8. In terms of covert attention, the average saliency map and mixelgram over time was computed. Thereby again only those frames contribute to the average that go along with speech from the parent. Also, the located maxima of each saliency and mutual information during parental speech are visualized. Here the maximum location within each frame contributes with a smooth spot to the visualization. Frequently attended locations appear brighter since the spots are overlaid.

The first video shows the demonstration of cup stacking (Fig. 7). A high amount of mutual information is concentrated on the face but also on the shirt. The richly structured texture of the pullover causes high activity in the image features due to minimal body movements. The highest average saliency is primarily found at the shoulders, where the subjects pullover sharply contrasts the background. In most frames, the maximum mutual information is located on the subject's face, whereas the maximum saliency is mostly found in the subjects action space (in the vicinity of the tutor's hands). Some exemplary frames are shown in Fig. 7(a). In some frames, the global mutual information maximum is located directly on the cups shown to the infant. Obviously, the cups are no source of sound in these situations but provide synchrony due to the interplay of parents' speech and motion.

The second video shows a demonstration of the wooden bricks (Fig. 8). On average, both saliency and mutual information show high values on the face and the right hands action space. The maximum mutual information is mostly found in the action space and—contrary to the first video—less often on the face. However, the synchrony is, in some frames, distracted towards, e.g., the infant's head, moving into the camera view. Also the saliency maxima are mainly restricted to this action space but not exclusively to the hands and objects as a maximum can, for instance, be found on the shirt's sticker. Generally, saliency is, however, less attracted by the subjects clothing compared to the first discussed video.

Taking both videos into account, one has to note that a perfect detection of task-relevant locations can be expected neither from synchrony or saliency nor from any bottom-up attention strategy. However, we can state that synchrony quite often points toward those locations and is hardly vulnerable to conspicuous modality-specific stimuli like textures or colors, where saliency maps are by design.



(a) Exemplary frames with positions of the mutual information (red) and saliency (green) maximum.

(b) Avg. mixelgram during parent speech and maximum mutual information positions

(c) Avg. saliency during parent speech and maximum saliency positions

Fig. 8. Subject demonstrating wooden bricks towards an infant.

### C. Discussion

The results presented in this section give a clear indication that multimodal motherese indeed involves a higher synchronization between gestures (or generally movement) and speech. Though this effect was also described by Gogate *et al.*, it is remarkable that it can be detected even at signal level. Moreover, our results show the synchronization for the first time by means of an objective, gradual measurement—avoiding a human's binary decision whether a condition is synchronous or not. The finding that more synchrony is detected throughout all four investigated tasks strengthens this observation. Also the significant correlation between AA and AC conditions indicates that our signal-level synchrony method is able to uncover systematic communication schemes.

It has been argued that cues from child-directed communication help to guide attention towards important parts of either the speech signal or the visual scene [35], [36], [6]. The shown spatial distributions and example frames suggest that mutual information can indeed be used to find relevant image locations. Gogate *et al.* found that object motion is often used synchronously to a word label in multimodal motherese. Though the correlation analysis is performed on an entirely different level, this is consistent with the observation that high mutual information values can be found on shown objects during parental speech.

However, we cannot generalize our results to a general communication scheme, since our analysis is limited to two videos. Here future investigations could, e.g., continue on the scheme used in [18], counting occurrences of maximum mutual information on the face, hands, and demonstrated objects throughout an entire set of videos. In particular, the relation between mutual information cues and purely visual saliency cues should be investigated. If audiovisual synchrony truly is an *additional* meaningful cue in multimodal motherese, there should be significant statistical differences between saliency and mutual information maxima.

So far we did not analyze the temporal characteristics of synchrony. It was argued [37] that child-directed speech has the function to *arouse* and *guide* the infant's attention. Whereas our study focuses on the guidance, it is also likely that synchrony in multimodal motherese is used to arouse the infant's attention when the child is currently not attending to the parent or the task. In that case, an increased level of synchrony might be measurable. Both functions are highly plausible in the context of the intersensory redundancy hypothesis [9] as young infants have been shown to preferentially attend to synchronous stimuli.

## IV. Conclusion and Outlook

We started right away from the question of how the mapping between words and concepts in word learning of infants is achieved. Following the arguments of Tomasello and others, we argued that it is crucial to consider it not as a simple batch learning task but rather as an interactive, social process. In our approach, we targeted a developmentally early modeling of binding, rather than a signal level, but in the context of a social learning scenario. In this scenario, we analyzed parents' behavior in a contingent interaction with their children. The effect that parents behave differently when tutoring infants compared to other adults by providing information in a more structured way has been termed "multimodal motherese" [6].

The driving research question for this paper was whether, and how, synchrony cues provided by a tutor can be detected with a computational model at signal level. Here, we clearly showed that the proposed model is able to uncover these cues. Though Prince [29] already confirmed that the model is basically suited for modeling infants synchrony detection, our work is the first to analyze this kind of model in an interactive tutoring situation, thus placing it in the context of social learning.

In congruence with the findings by Gogate *et al.* [6], also our results suggest a stronger intentional aim of the tutor when interacting with infants. The observed increase of correlation in infant-directed tutoring compared to adult-directed tutoring underpins the awareness of the tutor about how to manipulate the learner's attention, namely, by exploiting the infant's preference for intersensory redundancy. In particular, by our study and model, the presence of "multimodal motherese" is verified directly on the audio-visual signal in adult–infant tutoring.

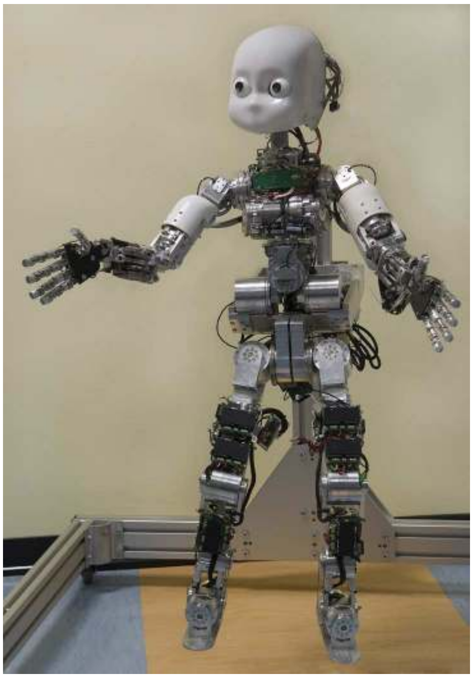


Fig. 9. The humanoid infant-like robot iCub.

### A. Application to Learning-Enabled Robots

From the very beginning of the conception of our algorithm, its applicability for artificial intelligent systems and its usability to enable them to learn has been in the focus of investigation. Consequently, one rationale for choosing particularly this computational model lies in its simplicity and efficiency. We particularly strive to endow robots—e.g., humanoids like iCub [38] (see Fig. 9)—with learning abilities resembling those of humans to a certain extent. In this sense, the work is a contribution to social robotics in general and to attentional models in robot learning in particular.

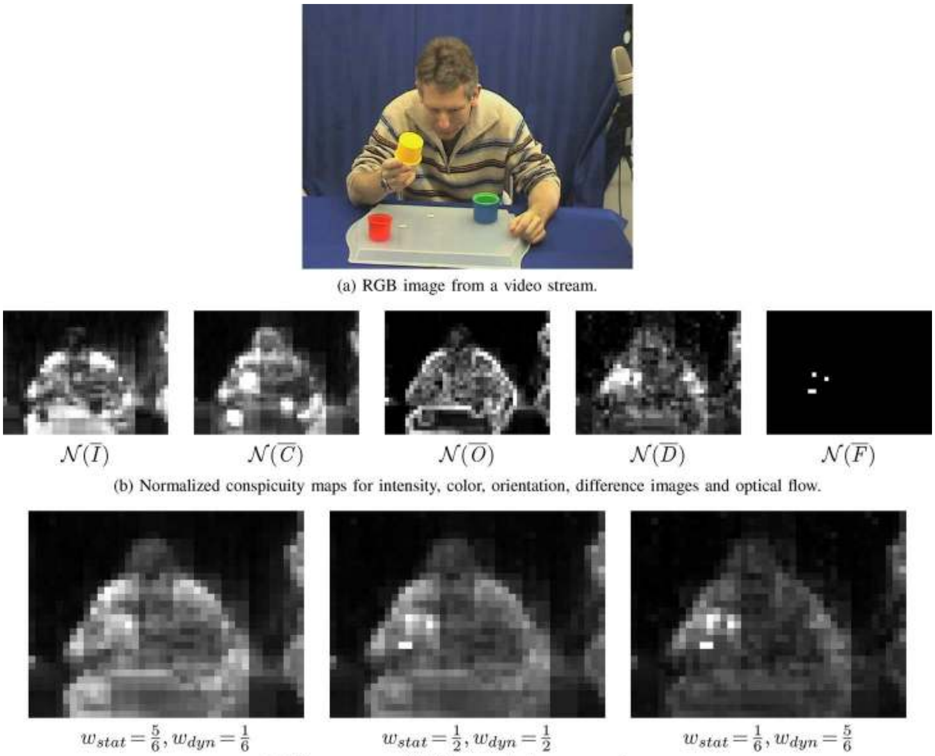
Specifically, interactive learning—also in robots—demands models of *joint* attention, in particular to assure an appropriate solution to the binding problem. Joint attention describes the effect that a social partner can influence the attention, which allows the interacting agents to simultaneously engage on one and the same external thing and form a kind of mental focus [39]. Kaplan and Haffner [40] distinguish four major requirements in order to achieve learning-enabling joint attention in robots: i) *attention detection* as the ability to understand attentional cues of the interaction partner; ii) *attention manipulation* relating to the ability to proactively affect the interaction partner's attentional behaviors; iii) *social coordination* controlling the actual course of interaction involving requirements such as turn-taking; and iv) *intentional stance* covering mutual context-aware interpretation and prediction. In the notion of Kaplan and Haffner, the presented computational model focuses on the detection of attention. It has been seconded by our findings that multimodal synchrony is a major cue for this detection.

For the challenge of attention detection, our computational model should, however, not be seen as an exclusive alternative to existing bottom-up attention models like saliency maps [33]. Rather, it provides a valuable addition and extension to them. Regarding subsequent processing, synchrony and saliency maps have the same structure and interpretation as they provide a topographic map of importance over the visual scene. Hence, it can be directly applied in any learning scenario exploiting visual saliency, such as presented by Nagai *et al.* [18].

An immediate integration of both concepts can, for instance, be accomplished via feature weighting schemes, which are also discussed as top-down strategy for visual search tasks [41], [42]. Since synchrony is always caused by the *dynamics* in a scene, more weight can be given to dynamic features (difference images, optical flow) in situations with high audiovisual synchrony. When there is no synchrony, the system can focus on static features like intensity, color, and orientation. In fact this view is consistent with the *intersensory redundancy hypothesis* [9]. The IRH claims that infants preferentially attend to amodal information in the presence of synchrony or redundancy across modalities. As amodal information, Bahrick *et al.* refer to, e.g., temporal, dynamic patterns. In the absence of redundancy across modalities, infants focus on modality-specific information like color.

Generally, this approach could be denoted with two weights $w_{\mathrm{stat}}(t)$ and $w_{\mathrm{dyn}}(t)$ for the static and dynamic features. The resulting saliency map is then a weighted sum of conspicuity maps, each conducted from a single feature. Increasing $w_{\mathrm{dyn}}(t)$—and therefore the relevance of the difference image and optical flow features—under synchronous/amodal stimulation forces a shift towards dynamic parts of the scene. An example of the possible impact of this approach is shown in Fig. 10.

The presented work fits in line with our general goal, which is to move away from batch learning in robotics towards "learning by interacting," as detailed in [43]. The assumption is that, similarly to infant-like learning, a lot of structuring is directly conveyed by interaction and that this structuring facilitates the learning also in robots. But up to now, most of the work regarding interactive social learning in robots—for instance, Leonardo [44] and BIRON [45]—only take already predefined interaction models as a means to facilitate learning. They build upon rather high-level interaction strategies and abilities such as gesture recognition and production, emotional modeling, and multimodal dialog management in order to achieve interactive tutoring. Hence, they regard objectives ii)-iv) identified by Kaplan and Haffner. Though these are also features of learning by interacting, they start at a higher level of cognitive abilities corresponding to later stages of development. For specific tasks and research questions, this is undoubtedly necessary; however, this paper was focused on the basal principles of learning in terms of the general binding problem to facilitate learning from scratch. With our focus on modeling the bottom-up pathway of multimodal attention in infant's learning, we lay a cornerstone to learning by interacting in robotics. Though for now we focused on an open-loop attentional model from the infant's perspective, it is assumed that it constitutes a prerequisite to any closed-loop interactive behaviors. In order to further analyze such closed-loop interactions, taking a closer look at the infant's and adult's responses to the multimodal tutoring is subject to ongoing investigations.


$\mathcal{N}(\overline{I})$
$\mathcal{N}(\overline{C})$
$\mathcal{N}(\overline{O})$
$\mathcal{N}(\overline{D})$
$\mathcal{N}(\overline{F})$
$w_{stat} = \frac{5}{6}, w_{dyn} = \frac{1}{6}$
$w_{stat} = \frac{1}{2}, w_{dyn} = \frac{1}{2}$
$w_{stat} = \frac{1}{6}, w_{dyn} = \frac{5}{6}$

(a) RGB image from a video stream.

(b) Normalized conspicuity maps for intensity, color, orientation, difference images and optical flow.

(c) Saliency maps with different weights $w_{stat}$ and $w_{dyn}$.

Fig. 10. Feature weighting based on a global synchrony measure: conspicuity maps are computed from a video stream. The bottom row shows different weights for the static features (intensity, color, and orientation) and dynamic features (difference images and optical flow). In situations without synchrony, high weight can be assigned to static features (left). In the presence of synchrony, the weighting promotes the dynamic features (right). A neutral, symmetric weighting is shown in the middle.

### B. Now, What is a “Gavagai”?

In fact, we did not try to answer this question but contributed to lines of research that investigate formation of multimodal binding to support, e.g., word learning but did not discuss the word learning itself. So, the question tackled here is more likely: “Which cues contribute to our knowledge about a gavagai?” In order to contribute to the complex space of answers to these questions, we focus on the analysis of multimodal synchrony, following the line of the intersensory redundancy hypothesis. In this view, intersensory redundancy is an important source of selective attention and learning in infancy [9]. A computational model correlating visual and auditory signals into a multimodal, spatial saliency model has been put forward as well for the offline analysis of adult-infant interactions as also for the online use as an attentional cue to facilitate learning in artificial intelligent systems. Relating our findings to the notion of synchrony of events as outlined in Section II, it can be stated that we presented a model to detect the on- and offsets of events implicitly present in the perceptual signals by synchrony. Thus, by means of these implicit event boundaries, information is structured and learning-relevant information is conveyed by tutors. Similar to preverbal infants, this information shall be exploited in order to enable robots to achieve a basal level of word-learning abilities following the developmental pathway, but in a social learning setting. Hence, one day our robot shall be enabled to answer what its own understanding of “gavagai“ is.

## REFERENCES

[1] W. Quine, *Word and Object*. Cambridge, MA: MIT Press, 1960.

[2] M. Tomasello, “Could we please lose the mapping metaphor, please?,” *Behav. Brain Sci.*, vol. 24, no. 6, pp. 1119–1120, 2001.

[3] C. Sinha, “Grounding, mapping and acts of meaning,” in *Cognitive Linguistics: Foundations, Scope and Methodology*, T. Janssen and G. Redeker, Eds. Berlin, Germany: Mouton de Gruyter, 1999, pp. 223–255.

[4] P. Zukow-Goldring, “Socio-perceptual bases for the emergence of language: An alternative to innatist approaches,” *Develop. Psychobiol.*, vol. 23, no. 7, pp. 705–726, 1990.

[5] P. Zukow-Goldring, “Assisted imitation: Affordances, effectivities, and the mirror system in early language development,” in *From Action to Language*, M. A. Arbib, Ed. Cambridge, U.K.: Cambridge Univ. Press, 2006, pp. 469–500.

[6] L. Gogate, L. Bahrick, and J. Watson, “A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures,” *Child Develop.*, vol. 71, no. 4, pp. 878–894, Jul./Aug. 2000.

[7] L. J. Gogate and L. E. Bahrick, “Intersensory redundancy and 7-month-old infants’ memory for arbitrary syllable-object relations,” *Infancy*, vol. 2, no. 2, pp. 219–231, 2001.

[8] L. J. Gogate, A. S. Walker-Andrews, and L. E. Bahrick, “The intersensory origins of word-comprehension: An ecological–dynamic systems view,” *Develop. Sci.*, vol. 4, no. 1, pp. 1–18, 2001.

[9] L. Bahrick, R. Lickliter, and R. Flom, “Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy,” *Current Direct. Psychol. Sci.*, 2004.

[10] J. Driver and C. Spence, “Multisensory perception: Beyond modularity and convergence,” *Current Biol.*, vol. 10, no. 20, pp. R731–R735, 2000.

[11] J. Vroomen and B. de Gelder, “Sound enhances visual perception: Cross-modal effects of auditory organization on vision,” *J. Exper. Psychol. Human Percept. Perform.*, vol. 26, no. 5, pp. 1583–1590, 2000.

[12] C. Binnie, A. Montgomery, and P. Jackson, “Auditory and visual contributions to the perception of consonants,” *J. Speech, Lang., Hearing Res.*, vol. 17, no. 4, pp. 619–630, 1974.

[13] B. Dodd, “The role of vision in the perception of speech,” *Perception*, vol. 6, no. 1, pp. 31–40, 1977.

[14] M. J. Mendelson and M. M. Haith, "The relation between audition and vision in the human newborn," *Mono. Soc. Res. Child Develop.*, vol. 41, no. 4, 1976.

[15] P. Kaplan, K. Fox, D. Scheuneman, and L. Jenkins, "Cross-modal facilitation of infant visual fixation: Temporal and intensity effects," *Infant Behavior Develop.*, vol. 14, no. 1, pp. 83–109, 1991.

[16] A. Meltzoff and P. Kuhl, "Faces and speech: Intermodal processing of biologically relevant signals in infants and adults," in *The Development of Intersensory Perception: Comparative Perspectives*, D. Lewkowicz and R. Lickliter, Eds. Philadelphia, PA: Lawrence Erlbaum, 1994.

[17] P. Zukow-Goldring, Rader, and N. de Villiers, "Caregiver gestures cultivate a shared understanding: Assisted imitation and early word learning," in *Intermodal Action Struct. Conf.*, Bielefeld, Germany, Jul. 2008.

[18] Y. Nagai and K. Rohlfing, "Can motionese tells infants and robots, "What to imitate"?," in *Proc. 4th Int. Symp. Imitation Animals Artifacts*, Apr. 2007.

[19] K. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?," *Adv. Robot.*, vol. 20, no. 10, pp. 1183–1199, 2006.

[20] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," *Nature Rev. Neurosc.*, vol. 5, pp. 495–501, 2004.

[21] D. J. Lewkowicz, "The development of intersensory temporal perception : An epigenetic systems/limitations view," *Psychol. Bull.*, vol. 126, no. 2, pp. 281–308, 2000.

[22] C. Spence and S. Squire, "Multisensory integration: Maintaining the perception of synchrony," *Current Biol.*, vol. 13, no. 13, pp. R519–R521, 2003.

[23] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 813–819, 2000.

[24] D. Matatyaho, Z. Mason, and L. Gogate, "Word learning by eight-month-old infants: The role of object motion and synchrony," in *Proc. 7th Int. Conf. Epigenetic Robot.*, 2007.

[25] V. Virsu, H. Oksanen-Hennah, A. Vedenpää, P. Jaatinen, and P. Lahti-Nuuttila, "Simultaneity learning in vision, audition, tactile sense and their cross-modal combinations," *Exper. Brain Res.*, vol. 186, no. 4, pp. 525–537, Apr. 2008.

[26] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, pp. 97–136, 1980.

[27] G. Monaci and P. Vandergheynst, "Audiovisual gestalts," in *Proc. 2006 Conf. Comput. Vision Pattern Recognit. Workshop*, 2006.

[28] H. Bredin and G. Chollet, "Measuring audio and visual speech synchrony: Methods and applications," in *Proc. IET Int. Conf. Visual Inf. Eng.*, 2006, pp. 255–260.

[29] C. G. Prince, G. J. Hollich, N. A. Helder, E. J. Mislivec, A. Reddy, S. Salunke, and N. Memon, "Taking synchrony seriously: A perceptual-level model of infant synchrony detection," in *Proc. 4th Int. Workshop Epigenetic Robot.*, 2004, pp. 89–96.

[30] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. NIPS*, 2000, pp. 814–820.

[31] E. Kidron, Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Process.*, vol. 55, pp. 1390–1404, 2005.

[32] C. Chibelushi, F. Deravi, and J. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, pp. 23–37, Mar. 2002.

[33] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[34] M. I. Posner, "Orienting of attention," *Quart. J. Exper. Psychol.*, vol. 32, no. 1, pp. 3–25, 1980.

[35] R. J. Brand, D. A. Baldwin, and L. A. Ashburn, "Evidence for 'motionese': Modifications in mothers' infant-directed action," *Develop. Sci.*, vol. 5, no. 1, pp. 72–83, 2002.

[36] P. F. Dominey and C. Dodane, "Indeterminacy in language acquisition: The role of child directed speech and joint attention," *J. Neurolinguist.*, vol. 17, no. 2-3, pp. 121–145, 2004.

[37] R. Cooper, J. Abraham, S. Berman, and M. Staska*, Infant Behavior Develop.*, vol. 20, no. 4, pp. 477–488, 1997.

[38] A. Cangelosi, T. Belpaeme, G. Sandini, G. Metta, L. Fadiga, G. Sagerer, K. Rohlfing, B. Wrede, S. Nolfi, D. Parisi, C. Nehaniv, K. Dautenhahn, J. Saunders, K. Fischer, J. Tani, and D. Roy, "The iTalk project: Integration and transfer of action and language knowledge," in *Proc. 3rd ACM/IEEE Int. Conf. Human Robot Interaction (HRI 2008)*, Amsterdam, The Netherlands, Mar. 2008.

[39] D. A. Baldwin, "Joint attention: Its origins and role in development," in *Understanding the Link Between Joint Attention and Language*. Philadelphia, PA: Lawrence Erlbaum, 1995, pp. 131–158.

[40] F. Kaplan and V. V. Hafner, "The challenges of joint attention," *Interact. Studies*, vol. 7, no. 35, pp. 135–169, 2006.

[41] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 3, no. 2, pp. 194–203, Mar. 2001.

[42] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychon. Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

[43] B. Wrede, K. J. Rohlfing, M. Hanheide, and G. Sagerer, "Towards learning by interacting," in *Creating Brain-Like Intelligence*, B. Sendhoff, E. Koerner, O. Sporns, H. Ritter, and K. Doya, Eds. Berlin, Heidelberg, Germany: Springer, 2009, pp. 139–150.

[44] A. Lockerd and C. Breazeal, "Tutelage and socially guided robot learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS 2004)*, 2004, vol. 4, pp. 3475–3480.

[45] J. Peltason, F. H. Siepmann, B. W. Thorsten, P. Spexard, M. Hanheide, and E. A. Topp, "Mixed-initiative in human augmented mapping," in *Proc. Int. Conf. Robot. Automat.*, 2009.

[46] P. G. Zukov, "Socio-perceptual basis for the emergence of language: An alternative to innatist approaches," *Develop. Psychobiol.*, vol. 23, no. 7, pp. 705–726, 1990.

**Matthias Rolf** received the Diploma degree (with distinction) in computer science from the Bielefeld University, Bielefeld, Germany, in 2008. The topic of his diploma thesis was the guidance of visual attention with audiovisual synchrony. Since 2008, he has been working towards the Ph.D. degree at the Research Institute for Cognition and Robotics, Bielefeld University, where his research topic is motor learning and control with neural networks.

His research interests are machine learning applications in robotics and vision, multimodal perception, developmental robotics, and computational neuroscience.

**Marc Hanheide** received the Diploma and the Ph.D. degree (Dr.-Ing.) in computer science from Bielefeld University, Bielefeld, Germany, in 2001 and 2006, respectively.

He worked in the European Union projects VAMPIRE (2005) and COGNIRON (2008). Currently, he is a Senior Researcher in the Applied Informatics Group and is responsible for several projects in cognitive interaction and robotics. His fields of research include human-robot interaction, multimodal perception, and intelligent systems. He is affiliated with the Research Institute for Cognition and Robotics (CoR-Lab) and the Cluster of Excellence Cognitive Interaction Technology (CITEC).

**Katharina J. Rohlfing** received the M.S. degree in linguistics, philosophy, and media studies from the University of Paderborn, Paderborn, Germany, in 1997. As a member of the Graduate Program "Task Oriented Communication," she received the Ph.D. degree (Dr. phil.) in linguistics from the Bielefeld University, Bielefeld, Germany, in 2002.

Her postdoctoral work at the San Diego State University, the University of Chicago, and Northwestern University was supported by a fellowship within the Postdoc program of the German Academic Exchange Service (DAAD) and by the German Research Foundation (DFG). In 2006, she became a Dilthey-Fellow (Funding initiative "Focus on the Humanities"), and her research is supported by the Volkswagen Foundation. Since May 2008, she has been Head of the Emergentist Semantics Group within Center of Excellence Cognitive Interaction Technology, Bielefeld University. She is interested in learning processes. In her research, she investigates the interface between early stages of language acquisition and cognitive development.