frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 05 August 2016
doi: 10.3389/frobt.2016.00046

Language Meddles with Infants' Processing of Observed Actions

Alessandra Sciutti[1], Katrin Solveig Lohan[2,3], Gustaf Gredebäck[4], Benjamin Koch[4] and Katharina J. Rohlfing[5]*

[1] Cognitive Interaction Lab, Robotics, Brain and Cognitive Sciences Department, Istituto Italiano di Tecnologia, Genova, Italy, [2] Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK, [3] Applied Informatics Group, Bielefeld University, Bielefeld, Germany, [4] Department of Psychology, Uppsala University, Uppsala, Sweden, [5] Faculty of Arts and Humanities, Psycholinguistics, Paderborn University, Paderborn, Germany

When learning from actions, language can be a crucial source to specify the learning content. Understanding its interactions with action processing is therefore fundamental when attempting to model the development of human learning to replicate it in artificial agents. From early childhood, two different processes participate in shaping infants' understanding of the events occurring around them: Infants' motor system influences their action perception, driving their attention to the action goal; additionally, parental language influences the way children parse what they observe into relevant units. To date, however, it has barely been investigated whether these two cognitive processes – action understanding and language – are separate and independent or whether language might interfere with the former. To address this question, we evaluated whether a verbal narrative concurrent with action observation could avert 14-month-old infants' attention from an agent's action goal, which is otherwise naturally selected when the action is performed by an agent. The infants observed movies of an actor reaching and transporting balls into a box. In three between-subject conditions, the reaching movement was accompanied either with no audio (Base condition), a sine-wave sound (Sound condition), or a speech sample (Speech condition). The results show that the presence of a speech sample underlining the movement phase reduced significantly the number of predictive gaze shifts to the goal compared to the other conditions. Our findings thus indicate that any modeling of the interaction between language and action processing will have to consider a potential top-down effect of the former, as language can be a meddler in the predictive behavior typical of the observation of goal-oriented actions.

Keywords: prediction, eye movement, development, action understanding, acoustic packaging, speech

Edited by:
Katerina Pastra,
Cognitive Systems Research Institute, Greece

Reviewed by:
Patric Bach,
Plymouth University, UK
Birgit Elsner,
University of Potsdam, Germany

****Correspondence:***
Alessandra Sciutti
alessanda.sciutti@iit.it

Specialty section:
This article was submitted to Computational Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *01 March 2016*
Accepted: *20 July 2016*
Published: *05 August 2016*

Citation:
Sciutti A, Lohan KS, Gredebäck G, Koch B and Rohlfing KJ (2016) Language Meddles with Infants' Processing of Observed Actions. Front. Robot. AI 3:46. doi: 10.3389/frobt.2016.00046

INTRODUCTION

From early childhood, humans demonstrate a marked ability in learning from demonstration. Any source of information, be it the observation of their parents' actions or their verbal instruction, are exploited to rapidly acquire new action and linguistic competencies. Enabling robots to learn to understand human actions as human infants do is a promising research line pursued in current robotics research (e.g., Rohlfing et al., 2006; Cangelosi et al., 2010; Ugur et al., 2015). However, it is not yet clear how infants integrate efficiently the multimodal tutoring inputs provided simultaneously by their parents, in order to understand their actions. Recently, Pastra (2013) reviewed robotic research suggesting language to be a spotlight or inducer of cognitive processes. In the current paper, we propose an additional dimension of how language can contribute to understanding actions by meddling with ongoing motor processes. Understanding this integration mechanism would be a key advancement for any attempt at replicating a similar learning process on a robotic platform.

With respect to ongoing motor processes, it has been suggested that the comprehension of others' action is substantially based on the involvement of the observer's own motor system, which allows to predict the agents' goals (Rizzolatti and Craighero, 2004; Falck-Ytter et al., 2006; Elsner et al., 2013; for a recent review, see Gredebäck and Falck-Ytter, 2015). Evidence in support to this mechanism comes for instance from the work by Gredebäck et al. (2009b), who have analyzed gaze behavior in infants as young as 10 and 14 months while they observed a continuous action sequence, in which a model reached for and displaced a series of objects from one location to another. The measured gaze patterns indicated that, already at 14 months of age, infants were able to segment the continuous action flow into sub-actions, by looking proactively at the goal of each sub-task [i.e., the object to be grasped or the target location of the transport – see also Baldwin et al. (2001) with a different methodology]. Recent research by Lakusta and Carey (2014) confirmed the limitation that only when the event involves action of an agent, 12-month-olds will give privilege to its goals. The results are in line with studies on adults suggesting that human action is perceived as hierarchically organized (Zacks and Tversky, 2001) with particular relevance given to action goal.

However, in today's approaches, researchers agree that language can function as a spotlight, making certain object properties highly salient even in non-linguistic thinking (Wolff and Holmes, 2010). For instance, Ferry et al. (2010) (but also Balaban and Waxman, 1997; Plunkett et al., 2008) have shown for 3-month-olds that words, but not other tones, highlight similarities between objects and facilitate categorization. The co-development of language and action was recognized in robotic research (Cangelosi et al., 2010) as an important mechanism supporting learning and representing compositional actions. For example, Schillingmann et al. (2009) have suggested that language can structure actions. Along these lines, Brand and Tapscott (2007) found that 9.5-month-old children consider sequences of actions that were "packaged" by a concurrent narration as belonging together. The general idea can be traced back to the work by Hirsh-Pasek and Golinkoff (1996) (p. 161), who proposed "acoustic packaging" as a means helping the child to find the boundaries of the event (ibid: p. 169), by segmenting the continuous visual input on the basis of the verbal signal. Indeed, Rolf et al. (2009) (also Schillingmann et al., 2009) have shown that the audio–visual coordination in children-directed interaction is greater than in adult-directed interaction. Gogate et al. (2000) tested mothers and their children at three different ages and found a greater audio–visual coordination for learning content (a new noun vs. non-target words) in interaction with younger infants aged 5–17 months. With more focus on what was said during action demonstrations, Meyer et al. (2011) have shown that maternal speech is synchronous to her actions when interacting with 6- to 13-month-old infants. More specifically, when demonstrating how to perform some actions, mothers seem to align speech and related action (rather than to align speech in general with actions), providing a meaningful multimodal behavior (see also Gogate and Bahrick, 2001). Hence, there is a strong argument for the role of language as social signal supporting cognitive processes in infancy such as categorization of objects and events. Therefore, language can also help infants to understand the actions they are looking at and to facilitate imitation (Southgate et al., 2009; Elsner and Pfeifer, 2012).

Thus, during early childhood, two different processes participate in shaping infants' understanding of the events occurring around them: on the one hand, infants' own motor development influences their perception driving their attention to the action goal (e.g., Falck-Ytter et al., 2006); on the other hand, parental language during interaction influences the way children parse what they observe into relevant units (e.g., Gogate et al., 2000). Our aim is to investigate how these two mechanisms, goal prediction and verbal input, interact in the development of action understanding in infancy. They sometimes cooperate, with parents' language principally emphasizing what infants are anticipating with their gaze (Lohan et al., 2014). However, they can also be in contrast, for instance when a tutor wants a child to focus not on the goal of the action she is performing, but on the way she is achieving that goal. In such a case, could language be an efficient tool to avert infants' attention from action goal, even without relying on infants' understanding of word semantics? And if so, in what way does language fulfill its function as a tool? One possibility is that it is processed as an additional cue, in which case infants would have to add the social signal on top of their event processing. Another possibility is that language influences directly the already ongoing event processing.

In this paper, we addressed these possibilities by evaluating gaze behavior during action observation in a group of 14-month-olds. We decided to test 14-month-olds, because Pruden et al. (2012) found only limited understanding of motion's manner in 12-month-olds. We presented participants with several reach-to-grasp actions, for which infants usually exhibit a clear goal anticipation with their gaze and evaluated whether a language stimulus could influence their attention. More precisely, we provided a narration overlapping with the movement, emphasizing the motion trajectory rather than the target object, and we evaluated whether such an acoustic signal could delay infants' gaze-shift to the goal.

We hypothesized that if infants perform a step-wise processing of the observed action, where language operates only after the occurrence of the motor-based mechanism that drives their attention to the goal, infants would discern the action's goal first and only subsequently they would be guided by the social signal toward the action's manner. Consistent with this view, children's attention could be attracted to parts other than action goal only by removing the landmark object (i.e., the action target Pruden et al., 2012) or modifying the action itself, by exaggerating the detail that needs to be attended to, cf. motionese (Brand et al., 2002; Rohlfing et al., 2006). An alternative hypothesis is that language can instead directly influence infants' action understanding, operating a top–down modification of the motor processing, even to the point of delaying their attentional shift to action goal.

MATERIALS AND METHODS

Subjects

Thirty-two 14-month-old infants (±4 days, 14 girls) learning Swedish and coming from Uppsala and its surroundings participated in the experiment; all were healthy and born within 2 weeks of expected date. This study was carried out in accordance with the recommendations of the Act concerning the Ethical Review of Research Involving Humans (2003, p. 460), Uppsala EPN, etikprövningsnämnderna i Uppsala. Parents provided written informed consent and received a gift certificate following participation. The study was conducted in accordance with the standards specified in the 1964 Declaration of Helsinki. Two subjects failed to provide any valid data point and were therefore excluded from further analysis.

Stimulus and Apparatus

Gaze was measured using a Tobii T120 near infrared eye tracker (Tobii, Sweden, Stockholm) with an infant add-on; precision 1°, accuracy 0.5°, and sampling rate 60 Hz. A standard nine-point infant calibration was used (Gredebäck et al., 2009a). The infants were seated at a distance of about 60 cm from a monitor (17″ size, 1024 × 768 pixel resolution) and were presented with movies (see Video S1 in Supplementary Material and **Figure 1A**). Each video lasted 20.73 s. The first action of the movie started with the hand of the demonstrator appearing from below the table about at the center of the scene after about 1.7 s, reaching for the first ball around second 3.7 and transporting it into the box at about second 6.5. This first sequence was followed by two reaching movements performed entirely in the fronto-parallel plane, from the box to the two remaining balls, alternated in a continuous manner with transport actions. The duration of the two reaching actions was of 2.73 and 2.86 s, respectively, while the corresponding transport actions lasted 2.87 and 2.86 s. We considered three different conditions (see **Figure 1A**): the Base condition, in which no audio was added to the movie, with the exception of the noise made by the ball falling into the container; the Sound condition, in which a 680 Hz sine-wave sound was added in correspondence of the reaching movements; and the Speech condition, where a verbal signal [a woman voice saying "Jag tar bollen" (I take the ball)] accompanied (or "packaged") the reaching movements (see in Video S1 in Supplementary Material the movies used as stimuli in the three conditions). The sound signal was played in correspondence of each reaching action in the Sound condition. It lasted 1.7 s in the first reaching, starting about 0.15 s after actor's hand appearance, and 2.3 s in the following two reaching actions, beginning about 0.35 s after the drop of the previous ball into the box (see **Figure 1B**). The verbal signal in the Speech condition was constituted by the word "Jag," which lasted around 200 ms and was followed, after about 160 ms,

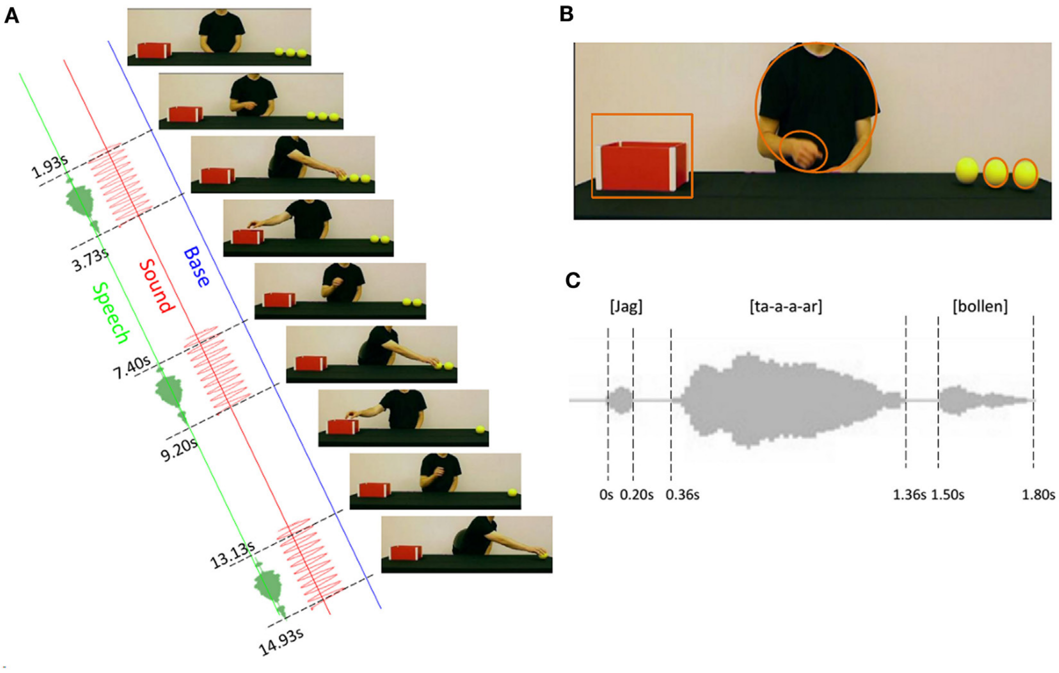

FIGURE 1 | Stimulus and procedure. Infants were presented with a video where an actor reached for three balls on a table and transported them into a box, in sequence. **(A)** A set of frames representing the reaching and transport of the balls with a schematic representation of the three experimental conditions. In the Base condition, no audio was associated to the video; in the Sound condition, a 680-Hz sine-wave sound was added in correspondence of the reaching movements and in the Speech condition a verbal signal accompanied the reaching movements (see Video S1 in Supplementary Material for a complete presentation of the stimuli used in the three conditions). **(B)** A frame of the video sequence with – superimposed – the five Areas of Interest adopted in the analysis of infants' gaze. **(C)** Visualization of the verbal signal associated to the reaching action in the Speech condition. The "tar" sound (corresponding to "take") was prosodically marked by extending the "a" vowel.

by the word "Tar" prosodically emphasized for about 1 s. The sentence was completed, after about 140 ms, by the word "Bollen," which lasted around 300ms, for a total duration of about 1.8 s (see **Figure 1C**). The timing of the speech signal was selected so that then end of the word "Bollen" was overlapped with the completion of the reaching action. Hence, the "Jag" utterance started about 0.26 s after hand appearance in the first reaching, 0.93 s after the drop of the first ball into the box for the second reaching and after 1.06 s after the drop of the second ball into the box for the third and last reaching.

Procedure

Each infant was seated on the parent's lap. After calibration, the presentation of the movie was replicated 10 times for each infant, interleaved with brief animations designed to orient infant's attention to the screen. Infants were divided into three groups, on the basis of the audio condition: the Base group (10 infants, 4 girls, 14.2 months), the Sound group (9 infants, 4 girls, 14.4 months), and the Speech group (11 infants, 4 girls, 14.4 months).

Data Reduction

Gaze positions collected by the Tobii gaze tracker were analyzed with custom made software for Icewing (Loemker, 2005). Five Areas Of Interest (AOI) were defined (**Figure 1B**): a circular one (Ball AOI, 2 visual degrees diameter) around each of the last two balls, a rectangular one over the box into which the balls were transported (Box AOI, 7.2° width, 3.2° height), an elliptic one around the actor's torso (longer axis, along the vertical direction: 19.2°, shorter axis: 17.5°), and another elliptic one around the actor's hand (Hand AOI, on average, longer axis: 3°, shorter axis: 2°). The sequence of hand positions was manually annotated beforehand, so the Hand AOI followed hand motion along its trajectory. The video was manually segmented into three reaching (to the balls) and three transport (to the box) sub-actions. As the future intention of an action sequence influences the degree of proactivity in each action component (Gredebäck et al., 2009b), we wanted to show to the infants the complete sequence of the actions used (i.e., reach-to-grasp-to-transport), before the real stimulus presentation. To this aim, the first reaching and transport of the each movie were used just to contextualize the actions and were not considered in the analysis. The analysis was carried out on the subsequent two movements in all the 10 video repetitions. For all subjects, the time difference between actor's hand and infant's gaze arrival on each ball AOI was computed for each reaching movement and the same difference with respect to the arrival to the Box AOI was measured for each transport motion. A positive value indicates that the infant looked at the target before the hand reached it. We chose to analyze (and "package" with the auditory stimuli) the reaching phases rather than the transport phases after a preliminary analysis of the data of the Base condition, which demonstrated that anticipation during the observation of the reaching actions was significantly larger than during the observation of the transport-into-the-box actions [two-sided, pair-sample *t*-test, $t(8) = 3.223$, $p < 0.012$, Cohen's $d = 1.52$], as suggested also by literature (Gredebäck et al., 2009b). All further analyses have been then conducted on the reaching portions of the videos. The gazing to the target was considered valid if, before reaching the ball AOI, the gaze moved to the actor's hand AOI. On average, 50% (±3% SD) of the trials complied with this constraint and were included in the analysis (49% for the Base condition, 49% for the Sound condition and 53% of the Speech condition). These values are in line with the subject-wise inclusion thresholds adopted in previous studies – e.g., minimum 26% of valid trials (Falck-Ytter et al., 2006). The measured time differences were aggregated into a single anticipation value for each subject. These data points were submitted to one-way ANOVAs followed by Tukey HSD *post hoc* tests for multiple comparisons, after checking normality with the Kolmogorov–Smirnov test and homogeneity of variance for the ANOVA through the Levene's test. Additionally, to evaluate whether infants got habituated during multiple stimulus presentations, we fitted linearly the anticipation measured on each trial for each group of infants as a function of trial number, after normalizing each subject's anticipations by dividing them by the value measured in the first valid trial, to cope with between subjects variability. For the infants in the Speech group we ran an additional analysis to assess whether the meaning of the verbal signal influenced their gaze patterns. In particular, we wanted to assess whether hearing "Jag" (=I) drove infants' attention to the actor's body and hearing "Bollen" (=ball), guided their gaze toward the object, suggesting an effect of the semantics of the signal. To this aim, we considered the time interval in which the word "Jag" was pronounced plus the following 300 ms (for a total of about 500 ms) and the time interval in which the word "Bollen" was uttered plus the subsequent 300 ms (for a total of about 600 ms). As a control time interval, we selected 500 ms in the middle of the "Tar" (=take) word, an interval equidistant in time from "Jag" and "Bollen" and in which the verbal signal did not highlight either the actor or the ball. For each of the selected intervals, we computed the percentage of gazing falling in the Actor area or at the Ball area over the total number of gazing executed in that interval. As Actor area, we considered both the Hand area and an area covering the torso of the actor (see **Figure 1B**). Pair-sample *t*-tests were performed to compare the average percentages of gazing to the Actor area during the "Jag" interval versus the control interval and the average percentages of gazing to the Ball area during the "Bollen" interval versus the control one.

RESULTS

To assess whether language associated to a reaching movement could modify infants' gazing pattern, we compared the timing of gaze arrival on target in the Base and the Speech conditions. Moreover, to be sure that the effect was specific for speech and did not just depend on the presence of a generic audio signal, we evaluated infants' anticipation of the target object also in a control condition where a sine-wave sound was associated to the video (Sound condition). Average anticipation in the Sound condition was of about 353 ms (±120 ms SEM), a value significantly different from 0 [$t(8) = 2.938$, $p = 0.0187$, Cohen's $d = 0.98$] and similar to the about 300 ms measured in the Base group. In the Speech condition, instead, the gaze arrived on action target about 34 ± 115 ms (SEM) after actor's hand arrival, i.e., −34 ms, a delay not significantly different from 0 [two-tailed, one-sample *t*-test,

$t(10) = -0.291$, $p = 0.777$, Cohen's $d = 0.88$]. A one-way ANOVA on anticipation with Condition as factor (three levels: Base, Sound and Speech) proved that the nature of the acoustic signal associated to the video can significantly influence infants' anticipatory behavior: $F(2, 27) = 3.78$, $p = 0.036$, partial $\eta^2 = 0.22$. The following one-tailed Tukey HSD *post hoc* tests showed that Speech selectively delayed infants' gaze shift to action goal ($p = 0.047$ and $p = 0.023$ for the comparison with Base and Sound respectively, while the Base-Sound comparison was not significant: $p = 0.462$, see **Figure 2A**).

The fact that the amount of anticipation is selectively reduced by the presence of a speech signal "packaging" the trajectory of the observed motion is confirmed also by an analysis of the percentage of presentations in which infants' gaze was anticipatory. In fact, a one-way ANOVA on percentages of anticipatory trials with Condition as factor individuated a significant impact of the auditory condition on gaze behavior [$F(2, 27) = 6.6305$, $p = 0.004$, partial $\eta^2 = 0.33$]. One-tailed Tukey HSD *post hoc* tests showed that no significant difference was present between the Sound and the Base groups ($p = 0.236$) while the percentage of anticipatory trials in those two conditions was significantly higher than that for Speech – Base: $p = 0.002$; Sound: $p = 0.039$, see **Figure 2B**).

Further, we checked whether habituation could occur during the repeated presentations of the movie. The subjects in the Base condition exhibited a significant decrease in the amount of anticipation, measured as the time difference between actor's hand and infant's gaze arrival on target, as a function of repetition number ($p = 0.043$), while habituation was not significant for the Sound and Speech conditions (p's = 0.086 and.536 respectively). To mitigate this effect, we repeated all the analyses considering just the first half of the presentations (first to fifth trials). All statistical tests reproduced similar results as the previous analysis, indicating a specific effect of the speech "packaging," which selectively reduced anticipation also during the earlier phase of the experiment.

To verify whether the effect of the speech signal was limited to the processing of the ongoing action or changed the overall strategy of action observation, we evaluated gaze behavior during the transport phases. The transports in fact followed the reaching sub-actions, but were not accompanied by any audio signal in all conditions. A one-way ANOVA did not reveal any significant difference in the timing of gaze arrival to the box as a function of the type of audio signal "packaging" the reaching phase: [$F(2, 26) = 0.166$, $p = 0.847$; −207 ± 78 ms (SEM), 174 ± 80 and 235 ± 70 ms after hand arrival for Base, Sound, and Speech, respectively]. This result confirms the hypothesis that speech selectively influenced the processing of the concurrent action, rather than modifying the way participants elaborate the whole action sequence.

Finally, to assess whether infants' gazing behavior in the Speech condition was affected by the meaning of the linguistic stimuli, we evaluated whether infants showed relatively more attention toward the actor during (and immediately after) the word "Jag" (I) and toward the target ball during (and immediately after) the word "Bollen" (Ball, see Materials and Methods). Infants' gaze did not seem to be substantially driven by the meaning of the words. In fact, the percentage of gazing to the actor was almost the same between the period of time around the "Jag" word and the control interval [2.9 ± 0.5% (SEM) and 3.1 ± 0.7% respectively, pair-sample t-test, $t(10) = 0.363$, $p = 0.724$, Cohen's $d = 0.16$] and between the period of time around the word "Bollen" and the control interval [10.7 ± 1.2% and 13.6 ± 1.6% respectively, pair-sample t-test, $t(10) = 1.887$, $p = 0.089$, Cohen's $d = 0.81$]. Therefore, no effect of word meaning was apparent in infants' gaze behavior, suggesting that semantics did not play a substantial role in determining attention allocation in our task.

DISCUSSION

The role of language as a social signal for cognitive development is crucial and has also been recognized in robotic research (Cangelosi et al., 2010). Several works have suggested that language can "educate" infants' attention, i.e., language have an impact on how events are presented and processed (Zukow-Goldring, 2006; Nomikou and Rohlfing, 2011). This means that by providing a steady social reinforcement in form of verbal behavior, infants might learn to cut out specific aspects of action and thus learn also subtle action differences. Thus, there are

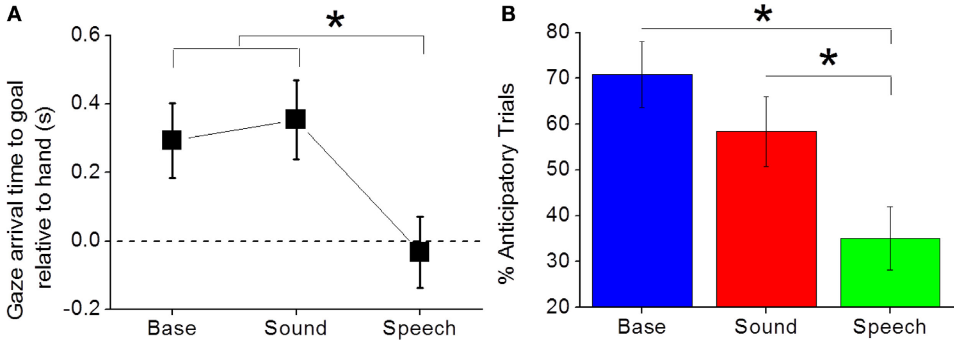

FIGURE 2 | **Anticipatory gaze is selectively reduced by acoustic packaging the trajectory of the observed motion**. **(A)** Anticipation, i.e., seconds between the arrival of the observer's gaze and the arrival of the actor's hand on the target ball in the three conditions. A positive number relates to anticipatory behavior. **(B)** Percentage of trials in which anticipation is larger than zero.

good reasons to assume that language guides attention to specific information about actions to infants. While since early infancy, humans are "obsessed" with goals, imitating, anticipating and looking at others' action goal, rather than to other aspects of the observed behaviors (Meltzoff, 1995; Woodward, 1999; Baldwin et al., 2001; Csibra and Gergely, 2007), language might be particularly important in driving infants' attention to other aspects of an action (e.g., its motion trajectory), possibly to facilitate his/her imitation learning of that specific characteristic (Southgate et al., 2009; Elsner and Pfeifer, 2012). In fact, studying actions Nagai and Rohlfing (2009) revealed that tutors structure their demonstrations providing some cues concerning the beginning or the end point. It was also found that during demonstrations, faces of the tutors were salient. Further research has shown that multimodal cues are used in child-directed interaction in order to manage attention when teaching not only where an object has to be moved but also *how* it has to be moved (Lohan et al., 2014). Therefore, a relevant question of our study was whether language has actually the power to meddle with infants' automatic attraction to action goal, allowing tutors to focus the infants' attention to other action properties.

We found that infants' anticipatory gaze shift to the goal of an observed action can be delayed by a verbal narrative concurrent with action presentation. The presence of a social signal in form of a sentence "Jag tar bollen [I take the ball]" underlining the movement phase of a reaching action reduced substantially the gaze proactivity usually exhibited by infants at this age. This phenomenon was specific for a social auditory stimulus whose structure was correlated with the action, because a simple sinus-wave tone, as provided in our control condition, was not able to induce any delay in the infants' shift of gaze to the target.

The fact that a concurrent narration can influence infants' perception of action units within a continuous action stream has been already demonstrated for 9.5-month-olds (see Brand and Tapscott, 2007). However, our findings extend the current evidence showing that in addition to an effect on action segmentation, verbal behavior can meddle with non-linguistic action processing, even downgrading the importance of a particular chunk of the action (the goal), which is commonly considered as more relevant by the child (Meltzoff, 1995; Woodward, 1998). Our results seem to indicate that language guides attentional processes during action presentation, even if this implies influencing an action processing mechanism which has been developed by the child prior to language understanding. Therefore, language can be considered as a meddler in the mechanisms associated to action execution and observation (Wolff and Holmes, 2010), exerting a top-down influence on the motor-based mechanism supporting action processing.

These findings are consistent with recent evidence on adults showing that language affects online action observation. In particular Hudson et al. (2016a) have demonstrated that hearing an actor declare his intention to take (or leave) an object before observing him reaching for (or withdrawing from) it, significantly modified the perceptual judgment of his hand position. In this study, however, the meaning of the verbal input played a relevant role, informing the observers of the actor's intention and biasing their perception even further toward the expected goal [see also Hudson et al. (2016b)]. Conversely, in our experiment infants did not seem to orient their attention on the basis of the meaning of the words they were listening at. Indeed, their looking patterns were not affected by the semantic of the words composing the verbal signals. On the contrary, language seemed to impact gaze behavior in force of its acoustic structure, averting the emphasis from the goal of the action. Our interpretation is that a verbal signal temporally and prosodically linked to the reaching action acted as a cue to a potential mismatch between the observed reaching and the observers' internal model. As a result, the voice interfered with the direct matching mechanism, causing a delay in the gaze shift toward action goal. This hypothesis seems to be confirmed by the observation that speech affected only the concurrent actions, with no significant impact on the subsequent, "unpackaged" transport movements.

It is worth noting that infants did not exhibit an anticipatory behavior during the observation of the transport actions, but rather tended to follow the actor's hand, even in the Base condition. This behavior is surprising, as anticipation is normally already present at 12 months of age when observing the transport of objects into a box (e.g., Falck-Ytter et al., 2006). The pattern of gaze timing is more similar to that measured for the observation of transport actions where multiple nearby target locations are present (cfr. Gredebäck et al., 2009b). An hypothesis is that the large dimension of the box and the slight variability of hand arrival location in our three transport actions might have increased the uncertainty about action goal, leading to this increase in latency, present only for this sub-part of the action sequence. Further studies will be required to verify this possibility.

Future research will need to analyze what property is responsible for the meddling effect of the Speech signal. Indeed, we did not test whether language had a particular relevance for its inherent social nature or if the auditory signal "structure" itself was sufficient to determine the observed modification in natural infants' gazing. Additional studies will be needed to disentangle the role of the prosody of the signal and that of its social nature to evaluate whether the acoustic packaging is tightly connected with the social (or emotional) aspects of the auditory input or if it can be extended also to artificial sounds.

To summarize, verbal stimuli can selectively modify infants' gaze patterns even if this implies meddling with a pre-existent action processing mechanism which would have shifted their attention toward action goal. Language becomes therefore a tool, through which parents can help their children to segment the continuous stream of observed actions into meaningful sub-components, also by modifying how they distribute their attention to the action. Such possibility could become relevant when parents want to teach the manner of an action, i.e., the specific movements needed to achieve a certain goal, which might become useful especially later on in development, when children will be faced with the need of learning complex tasks (as tying shoe laces) or specific movements for sports (as swimming or dancing).

From a computational perspective, our results have provided experimental evidence in favor of a direct interaction between different processes supporting cognition. In particular, the motor system, which is at the basis of the processing of observed actions and guides attention to action goal, is not immune to a top-down modulation by language. Any modeling effort aimed at building systems able to teach or learn from humans should take into account this potential interference. This finding has particular relevance for robotic research. Robots deployed in human environments will need to be able not only to interact in a socially appropriate manner but also to learn from their human partners. Learning gives robots the flexibility to adjust to the partner's specific needs and to perform actions in the context of the partner's environment. However, it is important that the learning process in the robot is not a cumbersome experience for the non-expert human tutor. A possible way to simplify the teaching process is to make robot learning as similar as possible to human learning, by making robots sensitive to the same stimulus properties that a human infant would notice (Rohlfing et al., 2006; Lohan et al., 2012). Indeed, it has been suggested that humans and robots have to process similarly the world they see to facilitate mutual understanding [e.g., finding the same types of stimuli salient (Breazeal and Scassellati, 1999)]. It becomes then relevant to understand in depth how the different processes involved in tutoring and learning interact. Drawing inspiration from the current human developmental study, we propose that future robots should start their learning by anticipating tutors' goal through action observation (Butz et al., 2003; Theofilis et al., 2013; Ugur et al., 2015). However this process should be overridable by any linguistic signal simultaneously provided by the tutor, averting robot's attention from the action target and potentially directing it toward *how* a goal is achieved. Action and language could then be processed together by the robot, leading to a seamless interaction with their tutor, who will be able to teach the machine using the same cognitive strategies adopted when interacting with children.

AUTHOR CONTRIBUTIONS

AS, KL, GG, BK, and KR gave substantial contributions to the conception of the work, the analysis and the interpretation of the data, and the drafting of the manuscript and its revision. All authors gave their final approval of the version to be published and agree to be accountable for all aspects of the work.

ACKNOWLEDGMENTS

This work has been supported by the European CODEFROR project (FP7-PIRSES-2013-612555) and by the EU project RobotDoC (235065) from the FP7 Marie Curie Actions ITN.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/frobt.2016.00046

Video S1 | **Experimental procedure, stimuli, and gaze patterns**. The video shows a schematic representation of the experimental setup, followed by the movies used as stimuli in the three conditions: Base, Sound, and Speech. The last portion of the movie shows representative gaze patterns for the Base (green dot) and the Speech (red dot) conditions, superimposed to the video stimulus. Each dot position at each frame has been computed as the median position for the Base and Speech groups of infants respectively, averaged further over the first five trials. Also from this qualitative analysis, it is evident that the verbal signal meddles with the low-level anticipatory gaze behavior, delaying infants' gaze shift to the goal of the reaching actions.

REFERENCES

Balaban, M. T., and Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *J. Exp. Child Psychol.* 64, 3–26. doi:10.1006/jecp.1996.2332

Baldwin, D. A., Baird, J. A., Saylor, M. M., and Clark, M. A. (2001). Infants parse dynamic action. *Child Dev.* 72, 708–717. doi:10.1111/1467-8624.00310

Brand, R. J., Baldwin, D. A., and Ashburn, L. A. (2002). Evidence for 'motionese': modifications in mothers' infant-directed action. *Dev. Sci.* 5, 72–83. doi:10.1111/1467-7687.00211

Brand, R. J., and Tapscott, S. (2007). Acoustic packaging of action sequences by infants. *Infancy* 11, 321–332. doi:10.1111/j.1532-7078.2007.tb00223.x

Breazeal, C., and Scassellati, B. (1999). "A context-dependent attention system for a social robot," in *International Joint Conference on Artificial Intelligence* (Stockholm).

Butz, M. V., Sigaud, O., and Gerard, P. (eds) (2003). "Anticipatory behavior in adaptive learning systems," in *Foundations, Theories, and Systems (LNAI 2684)* (Berlin: Springer), 1–10.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., et al. (2010). Integration of action and language knowledge: a roadmap for developmental robotics. *IEEE Trans. Auton. Ment. Dev.* 2, 167–195. doi:10.1109/TAMD.2010.2053034

Csibra, G., and Gergely, G. (2007). 'Obsessed with goals': functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychol (Amst)* 124, 60–78. doi:10.1016/j.actpsy.2006.09.007

Elsner, B., and Pfeifer, C. (2012). Movement or goal: goal salience and verbal cues affect preschoolers' imitation of action components. *J. Exp. Child Psychol.* 112, 283–295. doi:10.1016/j.jecp.2012.02.010

Elsner, C., D'Ausilio, A., Gredebäck, G., Falck-Ytter, T., and Fadiga, L. (2013). The motor cortex is causally related to predictive eye movements during action observation. *Neuropsychologia* 51, 488–492. doi:10.1016/j.neuropsychologia.2012.12.007

Falck-Ytter, T., Gredebäck, G., and von Hofsten, C. (2006). Infants predict other people's action goals. *Nat. Neurosci.* 9, 878–879. doi:10.1038/nn1729

Ferry, A. L., Hespos, S. J., and Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: an advantage of words over tones. *Child Dev.* 81, 472–479. doi:10.1111/j.1467-8624.2009.01408.x

Gogate, L. J., and Bahrick, L. E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy* 2, 219–231. doi:10.1207/S15327078IN0202_7

Gogate, L. J., Bahrick, L. E., and Watson, J. D. (2000). A study of multimodal motherese: the role of temporal synchrony between verbal labels and gestures. *Child Dev.* 71, 878–894. doi:10.1111/1467-8624.00197

Gredebäck, G., and Falck-Ytter, T. (2015). Eye movements during action observation. *Perspect. Psychol. Sci.* 10, 591–598. doi:10.1177/1745691615589103

Gredebäck, G., Johnson, S., and von Hofsten, C. (2009a). Eye tracking in infancy research. *Dev. Neuropsychol.* 35, 1–19. doi:10.1080/87565640903325758

Gredebäck, G., Stasiewicz, D., Falck-Ytter, T., Rosander, K., and von Hofsten, C. (2009b). Action type and goal type modulate goal-directed gaze shifts in 14-month-old infants. *Dev. Psychol.* 45, 1190–1194. doi:10.1037/a0015667

Hirsh-Pasek, K., and Golinkoff, R. M. (1996). *The Origins of Grammar: Evidence from Early Language Comprehension*. Cambridge, MA: MIT Press.

Hudson, M., Nicholson, T., Ellis, R., and Bach, P. (2016a). I see what you say: prior knowledge of other's goals automatically biases the perception of their actions. *Cognition* 146, 245–250. doi:10.1016/j.cognition.2015.09.021

Hudson, M., Nicholson, T., Simpson, W. A., Ellis, R., and Bach, P. (2016b). One step ahead: the perceived kinematics of others' actions are biased toward expected goals. *J. Exp. Psychol. Gen.* 145, 1–7. doi:10.1037/xge0000126

Lakusta, L., and Carey, S. (2014). Twelve-month-old infants' encoding of goal and source paths in agentive and non-agentive motion events. *Lang. Learn. Dev.* 11, 152–175. doi:10.1080/15475441.2014.896168

Loemker, F. (2005). *iceWing – A graphical Plugin Shell*. Available at: http://icewing.sourceforge.net/

Lohan, K. S., Griffiths, S. S., Sciutti, A., Partmann, T. C., and Rohlfing, K. J. (2014). Co-development of manner and path concepts in language, action, and eye-gaze behavior. *Top. Cogn. Sci.* 6, 492–512. doi:10.1111/tops.12098

Lohan, K. S., Rohlfing, K. J., Pitsch, K., Saunders, J., Lehmann, H., Nehaniv, C. L., et al. (2012). Tutor spotter: proposing a feature set and evaluating it in a robotic system. *Int. J. Soc. Robot.* 4, 131–146. doi:10.1007/s12369-011-0125-8

Meltzoff, A. N. (1995). Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Dev. Psychol.* 31, 1–16. doi:10.1037/0012-1649.31.5.838

Meyer, M., Hard, B., Brand, R. J., McGarvey, M., and Baldwin, D. A. (2011). Acoustic packaging: maternal speech and action synchrony. *IEEE Trans. Auton. Ment. Dev.* 3, 154–162. doi:10.1109/TAMD.2010.2103941

Nagai, Y., and Rohlfing, K. J. (2009). Computational analysis of motionese toward scaffolding robot action learning. *IEEE Trans. Auton. Ment. Dev.* 1, 44–54. doi:10.1109/TAMD.2009.2021090

Nomikou, I., and Rohlfing, K. J. (2011). Language does something. Body action and language in maternal input to three-month-olds. *IEEE Trans. Auton. Ment. Dev.* 3, 113–128. doi:10.1109/TAMD.2011.2140113

Pastra, K. (2013). Autonomous acquisition of sensorimotor experiences: any role for language? Newsletter on autonomous mental development. *IEEE Comput. Intell. Soc.* 10, 12–13. Available at: http://www.cse.msu.edu/amdtc/amdnl/

Plunkett, K., Hu, J. F., and Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition* 106, 665–681. doi:10.1016/j.cognition.2007.04.003

Pruden, S. M., Göksun, T., Roseberry, S., Hirsh-Pasek, K., and Golinkoff, R. M. (2012). Find your manners: how do infants detect the invariant manner of motion in dynamic events? *Child Dev.* 83, 977–991. doi:10.1111/j.1467-8624.2012.01737.x

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192. doi:10.1146/annurev.neuro.27.070203.144230

Rohlfing, K. J., Fritsch, J., Wrede, B., and Jungmann, T. (2006). How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Adv. Robot.* 20, 1183–1199. doi:10.1163/156855306778522532

Rolf, M., Hanheide, M., and Rohlfing, K. J. (2009). Attention via synchrony. Making use of multimodal cues in social learning. *IEEE Trans. Auton. Ment. Dev.* 1, 55–67. doi:10.1109/TAMD.2009.2021091

Schillingmann, L., Wrede, B., and Rohlfing, K. J. (2009). A computational model of acoustic packaging. *IEEE Trans. Auton. Ment. Dev.* 1, 226–237. doi:10.1109/TAMD.2009.2039135

Southgate, V., Chevallier, C., and Csibra, G. (2009). Sensitivity to communicative relevance tells young children what to imitate. *Dev. Sci.* 12, 1013–1019. doi:10.1111/j.1467-7687.2009.00861.x

Theofilis, K., Lohan, K. S., Nehaniv, C. L., Dautenhahn, K., and Werde, B. (2013). "Temporal emphasis for goal extraction in task demonstration to a humanoid robot by naive users," in *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)* (IEEE), 1–6.

Ugur, E., Nagai, Y., Sahin, E., and Oztop, E. (2015). Staged development of robot skills: behavior formation, affordance learning and imitation with motionese. *IEEE Trans. Auton. Ment. Dev.* 7, 119–139. doi:10.1109/TAMD.2015.2426192

Wolff, P., and Holmes, K. J. (2010). Linguistic relativity. *Wiley Interdiscip. Rev. Cogn. Sci.* 2, 253–265. doi:10.1002/wcs.104

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition* 69, 1–34. doi:10.1016/S0010-0277(98)00058-4

Woodward, A. L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behav. Dev.* 22, 145–160. doi:10.1016/S0163-6383(99)00007-7

Zacks, J. M., and Tversky, B. (2001). Event structure in perception and conception. *Psychol. Bull.* 127, 3–21. doi:10.1037/0033-2909.127.1.3

Zukow-Goldring, P. (2006). "Assisted imitation: affordances, effectivities, and the mirror system in early language development," in *From Action to Language*, ed. M. A. Arbib (Cambridge: CUP), 469–500.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.