frontiers in
NEUROROBOTICS

ORIGINAL RESEARCH ARTICLE
published: 21 May 2010
doi: 10.3389/fnbot.2010.00005

Integrating verbal and nonverbal communication in a dynamic neural field architecture for human–robot interaction

Estela Bicho[1], Luís Louro[1] and Wolfram Erlhagen[2]*

[1] *Department of Industrial Electronics, University of Minho, Guimarães, Portugal*
[2] *Department of Mathematics and Applications, University of Minho, Guimarães, Portugal*

Edited by:
Angelo Cangelosi, University of Plymouth, UK

Reviewed by:
Domenico Parisi, The National Research Council, Italy

****Correspondence:***
Wolfram Erlhagen, Department of Mathematics and Applications, University of Minho, 4800-058 Guimaraes, Portugal.
e-mail: wolfram.erlhagen@mct.uminho.pt

How do humans coordinate their intentions, goals and motor behaviors when performing joint action tasks? Recent experimental evidence suggests that resonance processes in the observer's motor system are crucially involved in our ability to understand actions of others', to infer their goals and even to comprehend their action-related language. In this paper, we present a control architecture for human–robot collaboration that exploits this close perception-action linkage as a means to achieve more natural and efficient communication grounded in sensorimotor experiences. The architecture is formalized by a coupled system of dynamic neural fields representing a distributed network of neural populations that encode in their activation patterns goals, actions and shared task knowledge. We validate the verbal and nonverbal communication skills of the robot in a joint assembly task in which the human–robot team has to construct toy objects from their components. The experiments focus on the robot's capacity to anticipate the user's needs and to detect and communicate unexpected events that may occur during joint task execution.

Keywords: joint action, neural fields, goal inference, natural communication, mirror system

INTRODUCTION

New generations of robotic systems are starting to share the same workspace with humans. They are supposed to play a beneficial role in the life of ordinary people by directly collaborating with them on common tasks. The role as co-worker and assistant in human environments leads to new challenges in the design process of robot behaviors (Fong et al., 2003). In order to guarantee user acceptance, the robot should be endowed with social and cognitive skills that makes the communication and interaction with the robot natural and efficient. Humans are experts in coordinating their actions with others to reach a shared goal (Sebanz et al., 2006). In collaborative tasks we continuously monitor the actions of our partners, interpret them effortlessly in terms of their outcomes and use these predictions to select an adequate complementary behavior. Think for instance about two people assembling a piece of furniture from its components. One person reaches toward a screw. The co-actor immediately grasps a screw-driver to hand it over and subsequently holds the components that are to be attached with the screw. In familiar tasks, such fluent team performance is very often achieved with little or no direct communication. Humans are very good in combining motion and contextual information to anticipate the ultimate goal of others' actions (Sebanz et al., 2006). Referring to objects or events through the use of language and communicative gestures is essential, however, whenever the observed behavior is ambiguous or a conflict in the alignment of intentions between partners has been detected. Ideally, not only the fact that something might go wrong in the joint action but also the reason for the conflict should be communicated to the co-actor.

The last decade has seen enormous progress in designing human-centered robots that are able to perceive, understand and use different modalities like speech, communicative gestures, facial expressions and/or eye gaze for more natural interactions with human users (for a recent overview see Schaal, 2007). Different control architectures for multi-modal communication have been proposed that address specific research topics in the domain of human-centered robotics. It has been shown for instance that integrating multiple information channels supports a more intuitive teaching within the learning by demonstration framework (McGuire et al., 2002; Steil et al., 2004; Pardowitz et al., 2007; Calinon and Billard, 2008), allows the robot to establish and maintain a face-to-face interaction in crowded environments (Spexard et al., 2007; Koenig et al., 2008), or can be exploited to guarantee a more intelligent and robust robot behavior in cooperative human–robot tasks (Breazeal et al., 2004; Alami et al., 2005; Foster et al., 2008; Gast et al., 2009). Although the proposed multi-modal architectures differ significantly in the type of control scheme applied (e.g., hybrid or deliberative) and theoretical frameworks used (e.g., neural networks, graphical or probabilistic models) they also have an important aspect in common. Typically, the integration of verbal and nonverbal information and the coordination of actions and decisions between robot and human are performed in dedicated fusion and planning modules that do not contain sensorimotor representations for the control of the robot actuators. A representative example are control architectures for HRI based on the theoretical framework of joint intention theory (e.g., Breazeal et al., 2004; Alami et al., 2005) that has been originally proposed for cooperative problem solving in distributed artificial intelligence systems (Cohen and Levesque, 1990). In these architectures a joint intention interpreter and a reasoner about beliefs and communicative acts can feed a central executive that is responsible for joint action planning and coordination on a symbolic level. A different approach to more natural and efficient HRI followed by our and

other groups is inspired by fundamental findings in behavioral and neurophysiological experiments analyzing perception and action in a social context (Wermter et al., 2004; Erlhagen et al., 2006b; Bicho et al., 2009; Breazeal et al., 2009). These findings suggest that automatic resonance processes in the observer's motor system are crucially involved in the ability to recognize and understand actions and communicative acts of others', to infer their goals and even to comprehend their action-related utterances. The basic idea is that people gain an embodied understanding of the observed person's behavior by internally simulating action consequences through the covert use of their own action repertoire (Barsalou et al., 2003). In joint action, the predicted sensory consequences of observed actions together with prior task knowledge may then directly drive the motor representation of an adequate complementary behavior. Such shared representations for perception, action and language are believed to constitute a neural substrate for the remarkable fluency of human joint action in familiar tasks (Sebanz et al., 2006).

Many of the experiments on action observation were inspired by the discovery of mirror neurons (MNs) first in premotor cortex and later in the parietal cortex of macaque monkey (di Pellegrino et al., 1992, for a review see Rizzolatti and Craighero, 2004). Mirror neurons fire both when the monkey executes an object-directed motor act like grasping and when it observes or hears a similar motor act performed by another individual. They constitute a neural substrate of an abstract concept of grasping, holding or placing that generalizes over agents and the modality of action-related sensory input. Many MNs require the observation of exactly the same action that they encode motorically in order to be triggered. The majority of MNs however falls in the broadly congruent category for which the match between observed and executed actions is not strict (e.g., independent of the kinematic parameters or the effector). Important for HRI, broadly congruent MNs may support an action understanding capacity across agents with very different embodiment and motor skills like human and robot. The fact that the full vision of an action is not necessary for eliciting a MN response whenever additional contextual cues may explain the meaning of the action has been interpreted as evidence for the important role of MNs in action understanding. It has been shown for instance that grasping MNs respond to a hand disappearing behind a screen when the monkey knew that there is an object behind the occluding surface (Umiltà et al., 2001). A grasping behavior is normally executed with an ultimate goal in mind. By training monkeys to perform different action sequences Fogassi et al. (2005) have recently tested whether MNs are not only involved in the coding of a proximate goal (the grasping) but also in the coding of the ultimate goal or motor intention (what to do with the object). The fundamental finding was that specific neural populations represent the identical grasping act in dependence of the outcome of the whole action sequence in which the grasping is embedded (e.g., grasping for placing versus grasping for eating). This finding has been interpreted as supporting the hypothesis that neural representations of motor primitives are organized in chains (e.g., reaching–grasping–placing) generating specific perceptual outcomes (Chersi et al., 2007, see also Erlhagen et al., 2007). On this view, the activation of a particular chain during action observation is a means to anticipate the associated outcomes of others' actions.

More recently, brain imaging studies of joint action revealed compelling evidence that the mirror system is also crucially involved in complementary action selection. People performing identical or complementary motor behaviors as those they had observed showed a stronger activation of the human mirror system in the complementary condition compared to the condition when the participants imitated the observed action (Newman-Norlund et al., 2007). This finding can be explained if one assumes a central role of the mirror system in linking two different but logically related actions that together constitute a goal-directed sequence involving two actors (e.g. receiving an object from a co-actor).

It has been suggested that the abstract semantic equivalence of actions encoded by MNs is related to aspects of linguistic communication (Rizzolatti and Arbib, 1998). Although the exact role of the mirror mechanism for the evolution of a full-blown syntax and computational semantics is still matter of debate (Arbib, 2005), there is now ample experimental evidence for motor resonance during verbal descriptions of actions. Language studies have shown that action words or action sentences automatically activate corresponding action representations in the motor system of the listener (Hauk et al., 2004; Aziz-Zadeh et al., 2006; Zwann and Taylor, 2006). Following the general idea of embodied simulation (Barsalou et al., 2003) this suggests that the comprehension of speech acts related to object-directed actions does not involve abstract mental representations but rather the activation of memorized sensorimotor experiences. The association between a grasping behavior or a communicative gesture like pointing and an arbitrary linguistic symbol may be learned when during practice the utterance and the matching hand movement occur correlated in time (Billard, 2002; Cangelosi, 2004; Sugita and Tani, 2005).

In this paper we present and validate a dynamic control architecture that exploits the idea of a close perception–action linkage as a means to endow a robot with nonverbal and verbal communication skills for natural and efficient HRI. Ultimately, the architecture implements a flexible mapping from an observed or simulated action of the co-actor onto a to-be-executed complementary behavior which consist of speech output and/or a goal-directed action. The mapping takes into account the inferred goal of the partner, shared task knowledge and contextual cues. In addition, an action monitoring system may detect a mismatch between predicted and perceived action outcomes. Its direct link to the motor representations of complementary behaviors guarantees the alignment of actions and decisions between the co-actors also in trials in which the human shows unexpected behavior.

The architecture is formalized by a coupled system of dynamic neural fields (DNFs) representing a distributed network of local neural populations that encode in their activation patterns task-relevant information (Erlhagen and Bicho, 2006). Due to strong recurrent interactions within the local populations the patterns may become self-stabilized. Such attractor states of the field dynamics allow one to model cognitive capacities like decision making and working memory necessary to implement complex joint action behavior that goes beyond a simple input–output mapping. To validate the architecture we have used a joint assembly task in which the robot has to construct together with a user different toy objects from their components. Different to our previous study in a symmetric construction task (Bicho et al., 2008, 2009), the robot does not directly participate in the construction work. The focus of the

present study is on anticipating the needs of the user (e.g., handing over pieces the user will need next) and on the detection and communication of unexpected events that may occur on the plan and the execution level. The robot reasons aloud to indicate in conjunction with hand gestures the outcome of its action simulation or action monitoring to the user. The robot is able to react to speech input confirming or not the prediction of the internal simulation process. It also understands object-directed speech commands (e.g., *Give me object X*) through motor simulation. The results show that the integration of verbal and nonverbal communication greatly improves the fluency and success of the team performance.

JOINT CONSTRUCTION TASK

For the human–robot experiments we modified a joint construction scenario introduced in our previous work (Bicho et al., 2009). The goal of the team is to assemble different toy objects from a set of components (**Figure 1**). Since these components are initially distributed in the separate working areas of the two teammates, the coordination of their actions in space and time is necessary in order to successfully achieve the task. The human performs the assembly steps following a given plan which explains the way how different pieces have to be attached to each other. He or she can directly request from the robot a specific component by using speech commands (e.g., *Give me component X*) and/or communicative hand gestures (e.g., pointing, requesting). The role of the robot is to hand over pieces in response to such requests or in anticipation of the user's needs, to monitor the user's actions and to communicate potential conflicts and unexpected behaviors during task execution to the user. Conflicts may result from a mismatch between expected and perceived goal-directed actions either because the action should have been performed later (sequence error) or the action is not compatible with any of the available construction plans defining possible target objects (wrong component).

The fact that the robot does not perform assembly steps itself simplifies the task representation that the robot needs to serve the user (for a symmetric construction scenario see Bicho et al., 2009). What the robot has to memorize is the serial order of the use of the different components rather than a sequence of subgoals (e.g., attach components A and B in a specific way) that have to be achieved

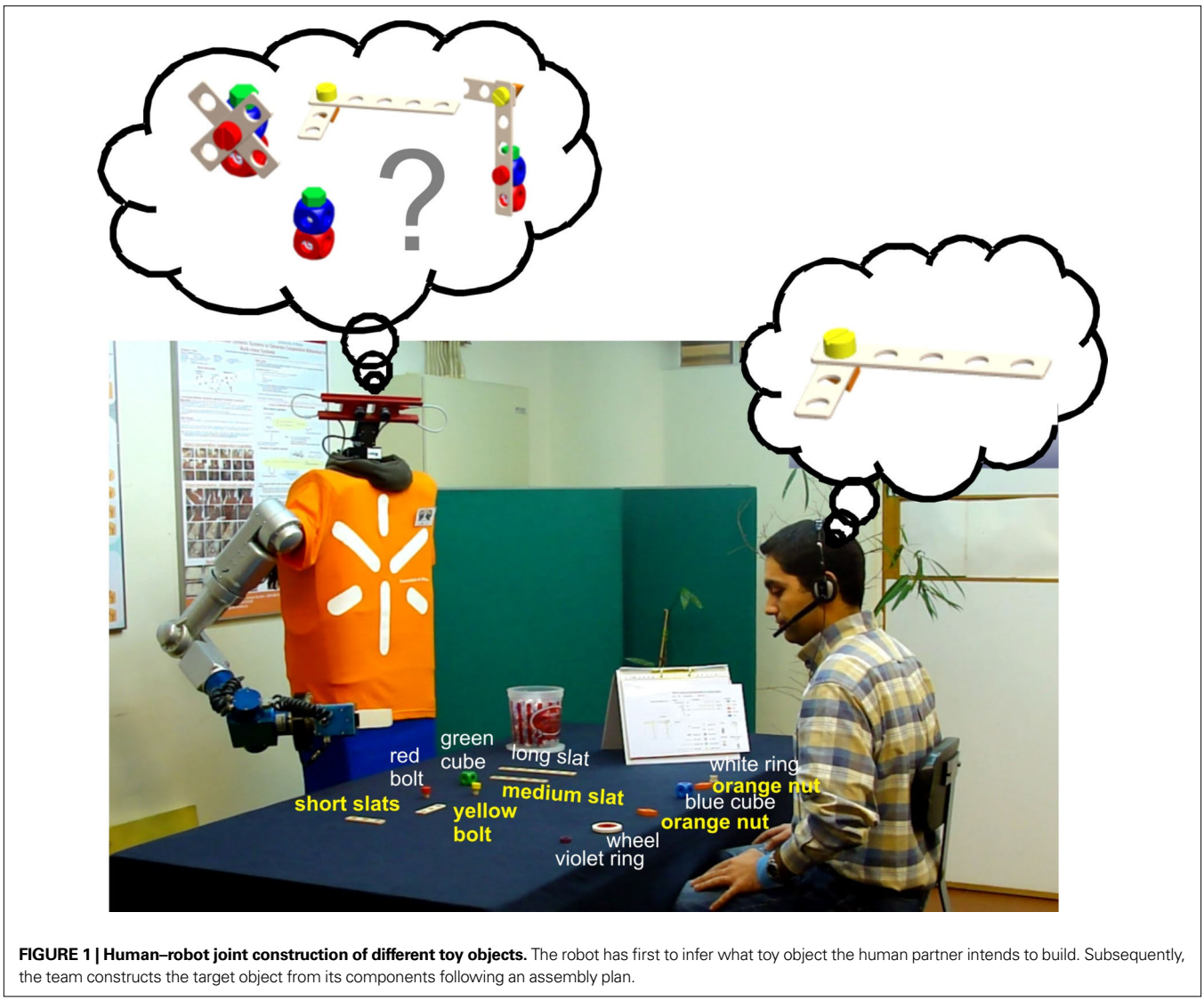

FIGURE 1 | Human–robot joint construction of different toy objects. The robot has first to infer what toy object the human partner intends to build. Subsequently, the team constructs the target object from its components following an assembly plan.

during the course of the assembly work. Importantly, since for each of the target objects the serial order of task execution is not unique, the robot has to simultaneously memorize several sequences of component-directed grasping actions in order to cope with different user preferences. To facilitate the coordination of actions and plans between the teammates, the robot speaks aloud and uses gestures to communicate the outcome of its goal inference and action monitoring processes to the user. For instance, the robot may respond to a request by saying *You have it there* and simultaneously points to the specific piece in the user's workspace. Although the integration of language and communicative gestures in the human–robot interactions will normally promote a more fluent task performance, this integration may also give rise to new types of conflict that the team has to resolve. From studies with humans it is well known for instance that if the verbally expressed meaning of an action or gesture does not match the accompanying hand movement (e.g., pointing to an object other than the object referred to) decision processes in the observer/listener appear to be delayed compared to a matching situation. This finding has been taken as direct evidence for the important role of motor representations in the comprehension of action-related language (Glenbach and Kaschak, 2002).

For the experiments we used the robot **ARoS** built in our lab. It consists of a stationary torus on which a 7 DOFs AMTEC arm (Schunk GmbH) with a two finger gripper and a stereo camera head are mounted. A speech synthesizer/recognizer (Microsoft Speech SDK 5.1) allows the robot to verbally communicate with the user. The information about object type, position and pose is provided by the camera system. The object recognition combines color-based segmentation with template matching derived from earlier learning examples (Westphal et al., 2008). The same technique is also used for the classification of object-directed, static hand postures such as grasping and communicative gestures such as pointing or demanding an object. For the control of the arm-hand system we applied a global planning method in posture space that allows us to generate smooth and natural movements by integrating optimization principles obtained from experiments with humans (Costa e Silva et al., submitted).

ROBOT CONTROL ARCHITECTURE

The multistage control architecture reflects empirical findings accumulated in cognitive and neurophysiological research suggesting a joint hierarchical model of action execution and action observation (van Schie et al., 2006; Hamilton and Grafton, 2008, see also Wolpert et al., 2003 for a modeling approach). The basic idea is that motor resonance mechanism may support social interactions on different but closely coupled levels: an intention level, a level describing the immediate goals necessary to realize the intention, and the kinematics level defining the movements of actions in space and time (**Figure 2A**).

Efficient action coordination between individuals in cooperative tasks requires that each individual is able to anticipate goals and motor intentions underlying the partner's unfolding behavior. As discussed in the introduction, most MNs represent actions on an abstract level sensitive to goals and intentions. For a human–robot team this is of particular importance since it allows us to exploit the motor resonance mechanism across teammates with very different embodiment.

In the following we briefly describe the main functionalities of the layered control architecture for joint action. It is implemented as a distributed network of DNFs representing different reciprocally connected neural populations. In their activation patterns the pools encode action means, action goals and intentions (or their associated perceptual states), contextual cues and shared task information (c.f. 'Model Details' for details on DNFs). In the joint construction task the robot has first to realize which target object the user intends to build. When observing the user reaching toward a particular piece, the automatic simulation of a reach-to-grasp action allows the robot to predict future perceptual states linked to the reaching act. The immediate prediction that the user will hold the piece in his/her hand is associated with the representation of one or more target objects that contain this particular part. In case that there is a one-to-one match, the respective representation of the target object becomes fully activated. Otherwise the robot may ask for clarification (*Are you going to assemble object A or object B?*) or may wait until another goal-directed action of the user and the internal simulation of action effects disambiguate the situation.

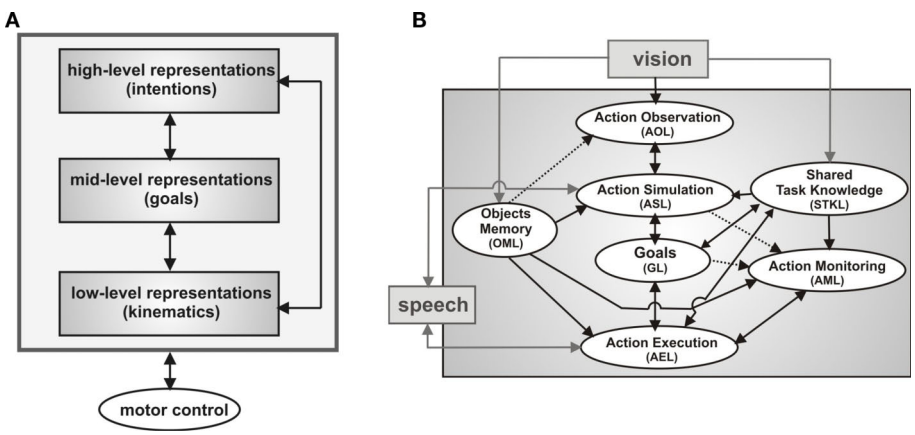

FIGURE 2 | Multistage robot control architecture. (A) Joint hierarchical model of action execution and action observation. **(B)** Mapping from observed actions (layer AOL) onto complementary actions (layer AEL) taking into account the inferred action goal of the partner (layer GL), detected errors (layer AML), contextual cues (layer OML) and shared task knowledge (layer STKL). The goal inference capacity is based on motor simulation (layer ASL).

Once the team has agreed on a specific target object, the alignment of goals and associated goal-directed actions between the teammates have to be controlled during joint task execution. **Figure 2B** presents a sketch of the highly context-sensitive mapping of observed onto executed actions implemented by the DNF-architecture. The three-layered architecture extends a previous model of the STS-PF-F5 mirror circuit of monkey (Erlhagen et al., 2006a) that is believed to represent the neural basis for a matching between the visual description of an action in area STS and its motor representation in area F5 (Rizzolatti and Craighero, 2004). This circuit supports a direct and automatic imitation of the observed action. Importantly for joint action, however, the model allows also for a flexible perception–action coupling by exploiting the existence of action chains in the middle layer PF that are linked to goal representations in prefrontal cortex. The automatic activation of a particular chain during action observation (e.g., reaching–grasping–placing) drives the connected representation of the co-actor's goal which in turn may bias the decision processes in layer F5 towards the selection of a complementary rather than an imitative action. Consistent with this model prediction, a specific class of MNs has been reported in F5 for which the effective observed and effective executed actions are logically related (e.g., implementing a matching between placing an object on the table and bringing the object to the mouth, di Pellegrino et al., 1992). For the robotics work we refer to the three layers of the matching system as the action observation (AOL), action simulation (ASL) and action execution layer (AEL), respectively. The integration of verbal communication in the architecture is represented by the fact that the internal simulation process in ASL may not only be activated by observed object-directed actions but also by action related speech input. Moreover, the set of complementary behaviors represented in AEL consists of goal-directed action sequences like holding out an object for the user but also contains communicative gestures (e.g., pointing) and speech output.

For an efficient team behavior, the selection of the most adequate complementary action should take into account not only the inferred goal of the partner (represented in GL) but also the working memory about the location of relevant parts in the separate working areas of the teammates (represented in OML), and shared knowledge about the sequential execution of the assembly task (represented in STKL). To guarantee proactive behavior of the robot, layer STKL is organized in two connected DNFs with representation of all relevant parts for the assembly work. Feedback from the vision system about the state of the construction and the observed or predicted current goal of the user will activate the population encoding the respective part in the first layer. Through synaptic links this activation pattern automatically drives the representations of one or more future components as possible goals in the second layer. Based on this information and in anticipation of the user's future needs the robot may already prepare the transfer of a part that is currently in its workspace.

In line with the reported findings in cognitive neuroscience the dynamic field architecture stresses that the perception of a co-actor's action may immediately and effortlessly guide behavior. However, even in familiar joint action tasks there are situations that require some level of cognitive control to override prepotent responses. For instance, even if the user would directly request verbally or by pointing a valid part located in the robot's workspace, the robot should not automatically start a handing over procedure. The user may have for instance overlooked that he has an identical object in his own working area. In this case, a more efficient complementary behavior for the team performance would be to use a pointing gesture to attract the user's attention to this fact. Different populations in the action monitoring layer (AML) are sensitive to a mismatch on the goal level (e.g., requesting a wrong part) or on the level of action means (e.g., handing over versus grasping directly). In the example, input from OML (representing the part in the user's workspace) and from ASL (representing the simulated action means) activate a specific neural population in AML that is in turn directly connected to the motor representation in AEL controlling the pointing gesture. As a result, two possible complementary actions, handing over and pointing, compete for expression in overt behavior. Normally, the pointing population has a computational advantage since the neural representations in AML evolve with a slightly faster time scale compared to the representations driving the handing over population. In the next section we explain in some more detail the mechanisms underlying decision making in DNFs. It is important to stress that the direct link between action monitoring and action execution avoids the problem of a coordination of reactive and deliberative components that in hybrid control architectures for HRI typically requires an intermediate layer (e.g., Spexard et al., 2007; Foster et al., 2008).

MODEL DETAILS

Dynamic neural fields provide a theoretical framework to endow artificial agents with cognitive capacities like memory, decision making or prediction based on sub-symbolic dynamic representations that are consistent with fundamental principles of cortical information processing. The basic units in DNF-models are local neural populations with strong recurrent interactions that cause nontrivial dynamic behavior of the population activity. Most importantly, population activity which is initiated by time-dependent external signals may become self-sustained in the absence of any external input. Such attractor states of the population dynamics are thought to be essential for organizing goal-directed behavior in complex dynamic situations since they allow the nervous system to compensate for temporally missing sensory information or to anticipate future environmental inputs.

The DNF-architecture for joint action thus constitutes a complex dynamical system in which activation patterns of neural populations in the various layers appear and disappear continuously in time as a consequence of input from connected populations and sources external to the network (e.g., vision, speech).

For the modeling we employed a particular form of a DNF first analyzed by Amari (1977). In each model layer i, the activity $u_i(x,t)$ at time t of a neuron at field location x is described by the following integro-differential equation (for mathematical details see Erlhagen and Bicho, 2006):

$$\tau_i \frac{\delta u_i(x,t)}{\delta t} = -u_i(x,t) + S_i(x,t) + \int w_i(x-x') f_i(u_i(x',t))dx' - h_i \quad (1)$$

where the parameters $\tau_i > 0$ and $h_i > 0$ define the time scale and the resting level of the field dynamics, respectively. The integral term describes the intra-field interactions which are chosen of lateral-inhibition type:

$$w_i(x) = A_i \exp\left(\frac{-x^2}{2\sigma_i^2}\right) - w_{\text{inhib},i} \quad (2)$$

where $A_i > 0$ and $\sigma_i > 0$ describe the amplitude and the standard deviation of a Gaussian, respectively. For simplicity, the inhibition is assumed to be constant, $w_{\text{inhib},i} > 0$. Only sufficiently activated neurons contribute to interaction. The threshold function $f_i(u)$ is chosen of sigmoidal shape with slope parameter β and threshold u_0:

$$f_i(u_i) = \frac{1}{1 + \exp[-\beta(u_i - u_0)]}. \quad (3)$$

The model parameters are adjusted to guarantee that the field dynamics is bi-stable (Amari, 1977), that is, the attractor state of a self-stabilized activation pattern coexists with a stable homogeneous activation distribution that represents the absence of specific information (resting level). If the summed input, $S_i(x,t)$, to a local population is sufficiently strong, the homogeneous state loses stability and a localized pattern in the dynamic field evolves. Weaker external signals lead to a subthreshold, input-driven activation pattern in which the contribution of the interactions is negligible. This preshaping by weak input brings populations closer to the threshold for triggering the self-sustaining interactions and thus biases the decision processes linked to behavior. Much like prior distributions in the Bayesian sense, multi-modal patterns of subthreshold activation may for instance model user preferences (e.g., preferred target object) or the probability of different complementary actions (Erlhagen and Bicho, 2006).

The existence of self-stabilized activation pattern allows us to implement a working memory function. Since multiple potential goals may exist and should be represented at the same time and all relevant components for the construction have to be memorized simultaneously, the field dynamics in the respective layers (STKL and ML) must support multi-peak solutions. Their existence can be ensured by choosing weight functions (Eq. 2) with limited spatial ranges. The principle of lateral inhibition can be exploited on the other hand to force and stabilize decisions whenever multiple hypothesis about the user's goal (ASL, GL) or adequate complementary actions (AEL) are supported by sensory or other evidence. The inhibitory interaction causes the suppression of activity below resting level in competing neural pools whenever a certain subpopulation becomes activated above threshold. The summed input from connected fields u_l is given as $S_i(x,t) = k\Sigma_l S_l(x,t)$. The parameter k scales the total input to a certain population relative to the threshold for triggering a self-sustained pattern. This guarantees that the inter-field couplings are weak compared to the recurrent interactions that dominate the field dynamics (for details see Erlhagen and Bicho, 2006). The scaling also ensures that missing or delayed input from one or more connected populations will lead to a subthreshold activity distribution only. The input from each connected field u_l is modeled by Gaussian functions:

$$S_l(x,t) = \sum_m \sum_j a_{mj} c_l(t) \exp\left(\frac{-(x - x_m)^2}{2\sigma^2}\right) \quad (4)$$

where $c_l(t)$ is a function that signals the presence or absence of a self-stabilized activation peak in u_l, and a_{mj} is the inter-field synaptic connection between subpopulation j in u_l to subpopulation m in u_i. Inputs from external sources (speech, vision) are also modeled as Gaussians for simplicity.

RESULTS

In the following we discuss results of real-time human–robot interactions in the joint construction scenario. The snapshots of video sequences shall illustrate the processing mechanisms underlying the robot's capacity to anticipate the user's need and to deal with unexpected events. To allow for a direct comparison between different joint action situations, the examples all show the team performance during the construction of a single target object called L-shape (**Figure 3**). Details on the connection scheme for the neural pools in the layered architecture and numerical values for the DNF parameters and inter-field synaptic weights may be found in the Supplementary Material.

The initial communication between the teammates that lead to the alignment of their intentions and plans is included in the videos. They can be found at http://dei-s1.dei.uminho.pt/pessoas/estela/JASTVideosFneurorobotics.htm. The plan describing how and in which serial order to assemble the different components is given to the user at the beginning of the trials. We focus the discussion of results on the ASL and AEL. **Figures 4, 5 and 7** illustrate the experimental results. In each Figure, panel A shows a sequence of video snapshots, panel B and C refer to the ASL and AEL, respectively. For both layers, the total input (top) and the field activation (bottom) are compared for the whole duration of the joint assembly work. **Tables 1 and 2** summarize the component-directed actions and communicative gestures that are represented by different populations in each of the two layers. Since the robot does not perform assembly steps itself, AEL only contains two types of overt motor behavior: pointing towards a specific component in the user's workspace or grasping a piece for holding it out for the user.

It is important to stress that the dynamic decision making process in AEL also works in more complex situations with a larger number of possible complementary action sequences linked to each component (Erlhagen and Bicho, 2006).

Figure 4 shows the first example in which the humans starts the assembly work by asking for a medium slat (S1). The initial distribution of components in the two workplaces can be seen in **Figure 1**. The fact that the user simultaneously points towards a short slat creates a conflict that is represented in the bi-modal input pattern to ASL centered over A6 and A7 at time T0. As can be seen in the bottom layer of **Figure 4B**, the field dynamics of ASL resolves this conflict by evolving a self-sustained activation pattern. It represents a simulated pointing act towards the short slat. The decision is the result of a slight difference in input strength which favors communicative gestures over verbal statements. This bias can be seen as reflecting an interaction history with different users. Our human–robot experiments revealed that naive users are usually better in pointing than verbally referring to (unfamiliar) objects. The robot directly communicates the inferred goal to the

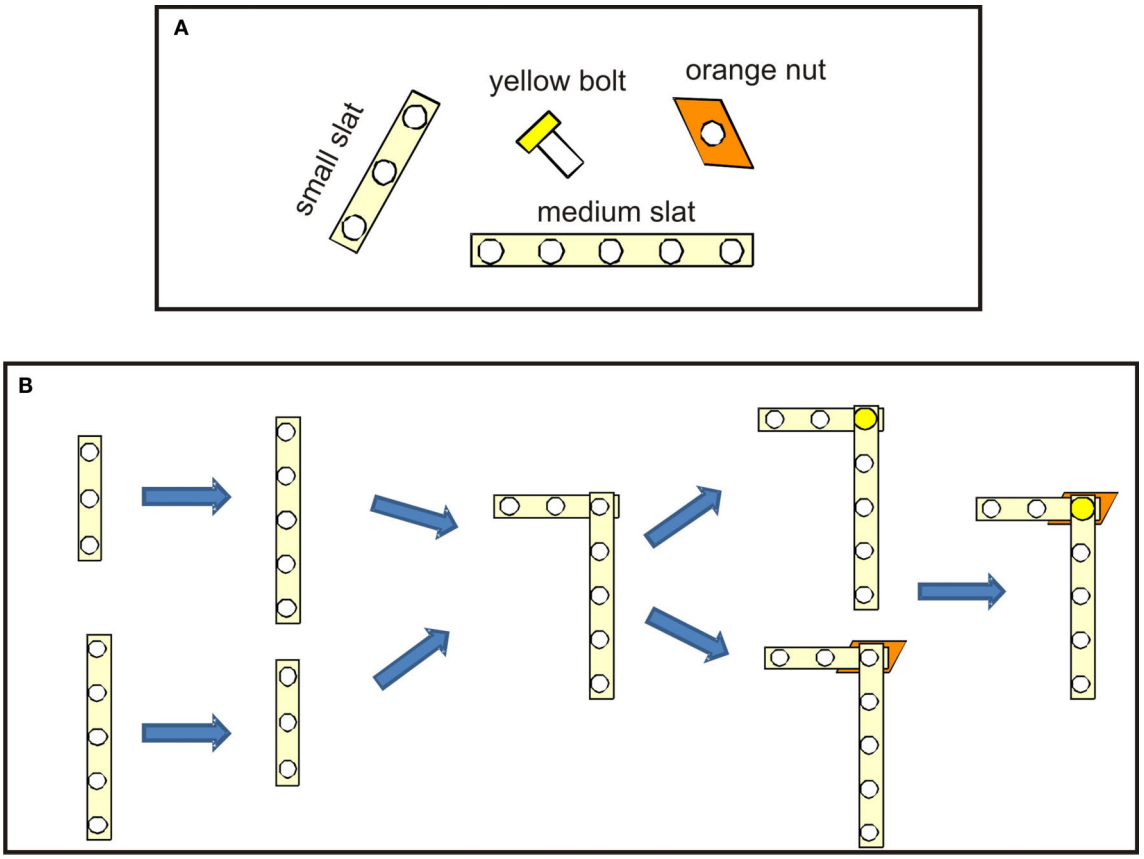

FIGURE 3 | Toy object L-shape. (A) Pieces used to build the L-shape. **(B)** Different serial orders to assemble the L-shape.

user (S2). **Figure 4C** shows that the input to AEL supports two different complementary actions, A1 and A2. However, since the total input from connected layers is stronger for alternative A1, the robot decides to hand over the short slat (S3). Subsequently, the robot interprets the user's request gesture (empty hand, S4) as demanding a medium slat (S5). The observed unspecific gesture activates to some extent all motor representations in ASL linked to components of the L-shape in the robot's workspace (compare the input layer). Goal inference is nevertheless possible due to the input from STKL that contains populations encoding the sequential order of task execution. The field activation of AEL (**Figure 4C**) shows at time T1 the evolution of an activation peak representing the decision to give the medium slat to the user (S6). At time T2 the robot observes the human reaching towards an orange nut (S7). The visual input from AOL activates the motor representation A4 in ASL which enables the robot to predict that the human is going to grasp the nut (S7). Since according to the plan the nut is followed by a yellow bolt and the bolt is in its workspace, the robot immediately starts to prepare the handing over procedure and communicates the anticipated need to the user (S8–S9). Note that the activation patterns representing the inferred current goal of the user (A4 in ASL) and the complementary action (A3 in AEL) evolve nearly simultaneously in time. An additional observation is worth mentioning. The input supporting the complementary behavior A3 starts to increase shortly after the decision to hand over the medium slat, that is, well ahead of the time when the robot predicts the nut as the user's next goal. This early preparation reflects the fact that handing over the medium slat automatically activates the representations of all possible future goals in STKL that are compatible with stored sequential orders. Since a yellow bolt and an orange nut represent both possible next assembly steps, the combined input from STKL and OML (bolt in robot's workspace) explains this early onset of subthreshold motor preparation in AEL.

In the second example (**Figure 5**) the initial distribution of components in the two working areas is identical to the situation in the first example. However, this time the meaning of the verbal request and the pointing act are congruent. Consequently, the input converges on the motor representation in ASL representing the pointing (A6) and a suprathreshold activity pattern quickly evolves. This in turn activates the population encoding the complementary behavior of handing over the short slat in AEL. Compared to the dynamics of the input and the field activity in the previous case (**Figure 4C**) one can clearly see that in the congruent condition the input arrives earlier in time and the decision process is faster. Note that in both cases the alternative complementary behavior representing the transfer of a medium slat (A3) appears to be activated below threshold at time T0. This pre-activation is caused by the input from STKL that supports both the short and the medium slat as possible goals at the beginning of the assembly work.

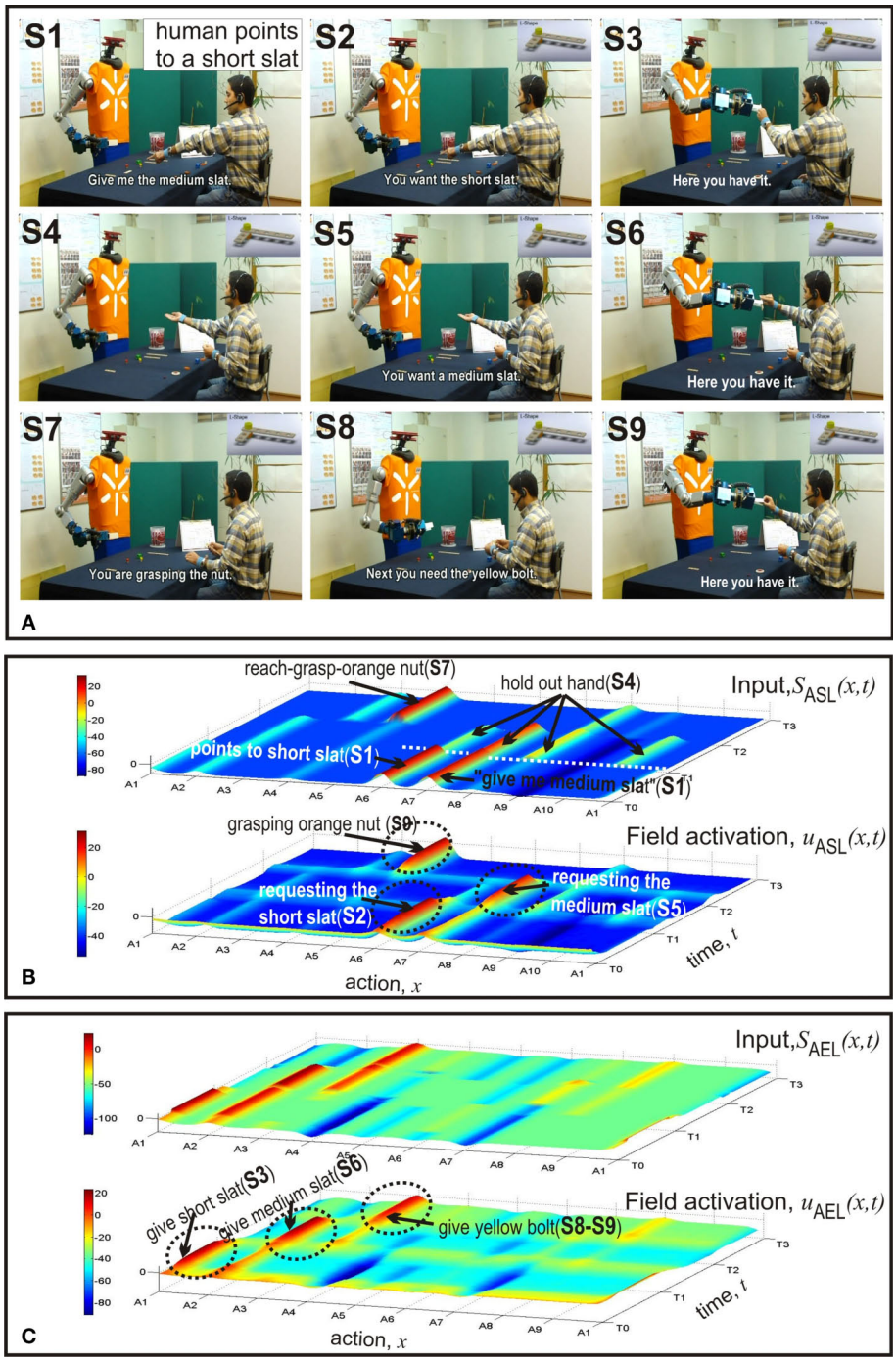

FIGURE 4 | First example: (1) goal inference when gesture and speech contain incongruent information (ASL), and (2) anticipatory action selection (AEL). (A) Video snapshots. **(B)** Temporal evolutions of input to ASL (top) and activity in ASL (bottom). **(C)** Temporal evolutions of input to AEL (top) and activity in AEL (bottom).

In the third example (**Figures 6 and 7**) the robot's action monitoring system detects a sequence error and the robot reacts in an appropriate manner before the failure becomes manifested. The robot observes a reaching towards the short slat (S1) and communicates to the user that it infers the short slat as the user's goal (S2). The input to the AEL (**Figure 7C**) triggers at time T0 the evolution of an activation pattern at A6 representing the preparation of a pointing to the medium slat in the user's workspace. However, this pattern does not become suprathreshold since at time T1 the user request the yellow bolt in the robot's workspace (S3). By internally simulating a pointing gesture the robot understands the request (S4) which in turn causes an activity burst of the population in AEL representing the corresponding complementary behavior (A3). However, also this pattern does not reach the decision level due to

Table 1 | Goal-directed sequences and communicative gestures in ASL.

Action	Sequence of motor primitives	Short description
A_1	Reach short slat → grasp	Use short slat
A_2	Reach medium slat → grasp	Use medium slat
A_3	Reach yellow bolt → grasp	Use yellow bolt
A_4	Reach orange nut → grasp	Use orange nut
A_5	Reach other piece → grasp	Use other part
A_6	Point to short slat	Request short slat
A_7	Point to medium slat	Request medium slat
A_8	Point to yellow bolt	Request yellow bolt
A_9	Point to orange nut	Request orange nut
A_{10}	Point to other part	Request other part

Table 2 | Goal-directed sequences and communicative gestures in AEL.

Action	Sequence of motor primitives	Short description
A_1	Reach short slat → grasp	Give short slat
A_2	Reach medium slat → grasp	Give medium slat
A_3	Reach yellow bolt → grasp	Give yellow bolt
A_4	Reach orange nut → grasp	Give orange nut
A_5	Point to short slat	Attend to short slat
A_6	Point to medium slat	Attend to medium slat
A_7	Point to yellow bolt	Attend to yellow bolt
A_8	Point to orange nut	Attend to orange nut
A_9	Point to other part	Attend to other part

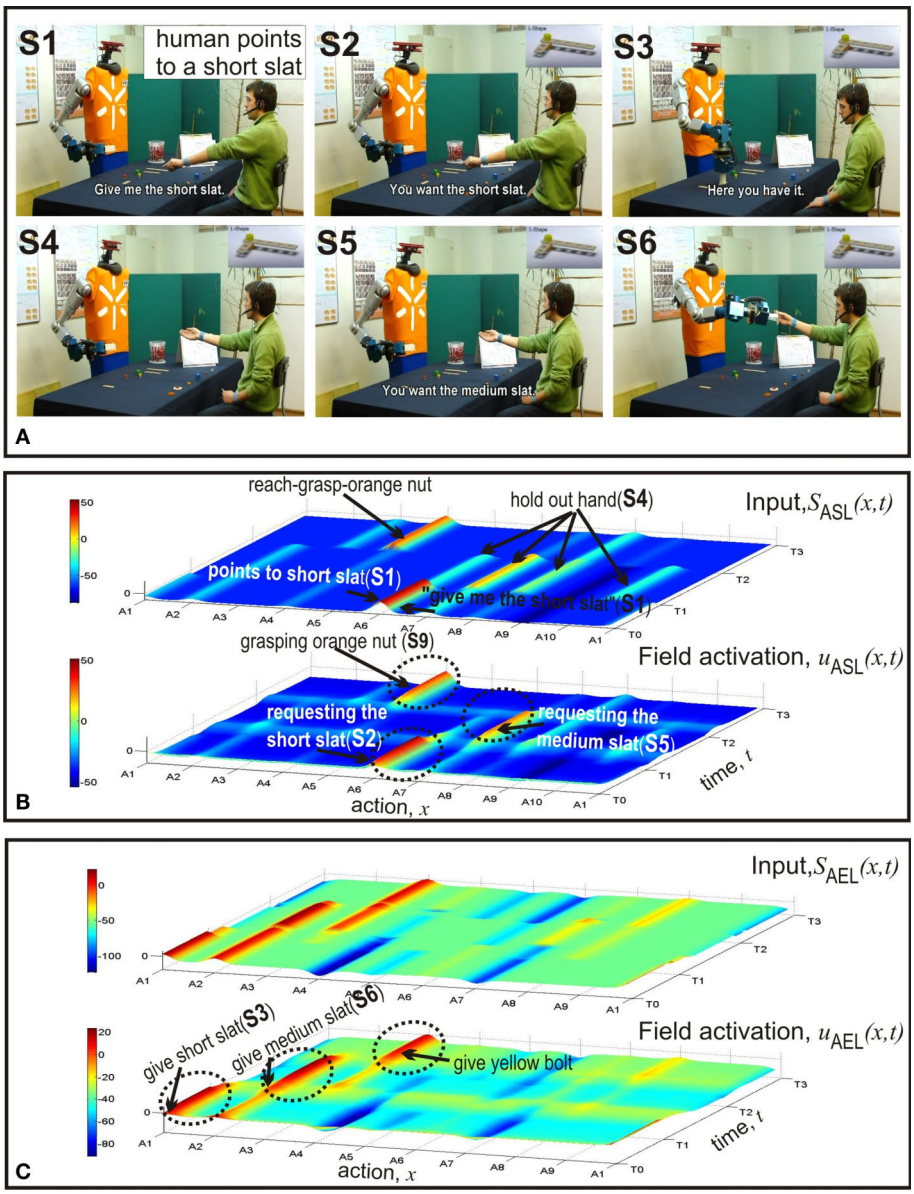

FIGURE 5 | Second example: faster goal inference and speeded decision making due to congruent information from gesture and speech. (A) Video snapshots. **(B)** Temporal evolutions of input to ASL (top) and activity in ASL (bottom). **(C)** Temporal evolutions of input to AEL (top) and activity in AEL (bottom).

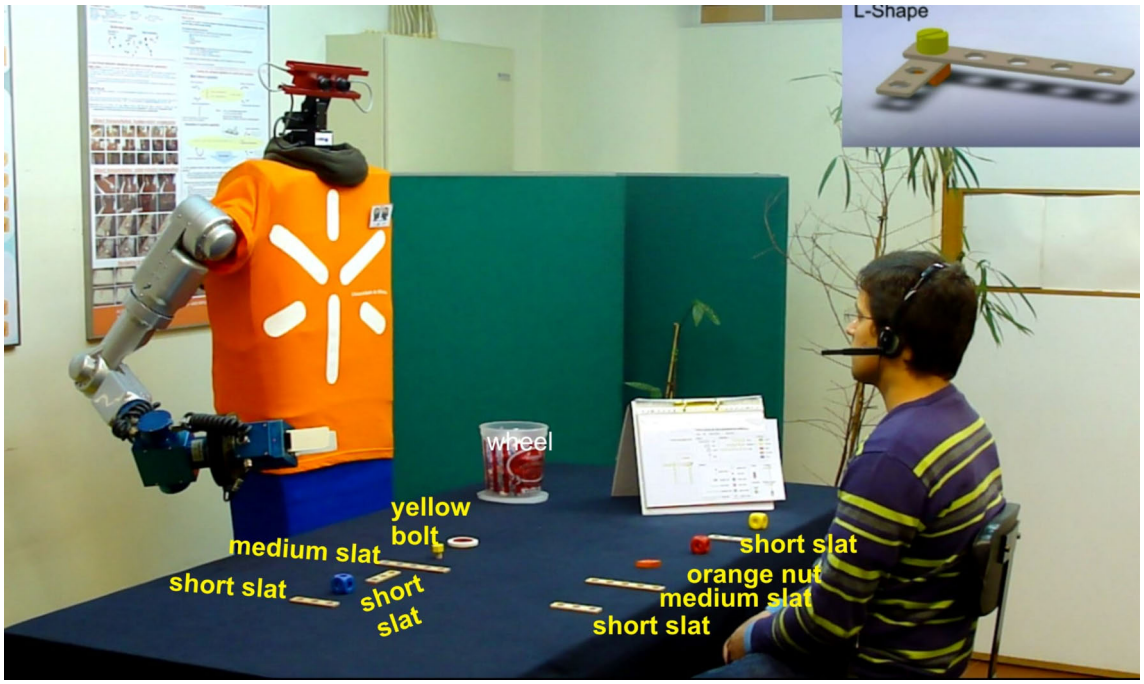

FIGURE 6 | Third example: initial distribution of components in the two working areas.

inhibitory input from a population in the AML. This population integrates the conflicting information from STKL (possible goals) and the input from the action simulation (yellow bolt). The robot informs the user about the sequence error (S5) and suggests the correction by pointing towards the medium slat and speaking to the user (S6). The pointing gesture is triggered by converging input from STKL, OML and the population in AML representing the conflict. The user reacts by reaching towards the correct piece (S7). The internal simulation of this action triggers the updating of the goals in STKL which allows the robot to anticipate what component the user will need next. As shown by the suprathreshold activation pattern of population A3 in AEL, the robot immediately prepares the transfer of the yellow bolt (S8–S9).

DISCUSSION AND SUMMARY

The main aim of the present study was to experimentally test the hypothesis that shared circuits for the processing of perception, action and action-related language may lead to more efficient and natural human–robot interaction. Humans are remarkably skilled in coordinating their own behavior with the behavior of others to achieve common goals. In known tasks, fluent action coordination and alignment of goals may occur in the absence of a full-blown human conscious awareness (Hassin et al., 2005). The proposed DNF-architecture for HRI is deeply inspired by converging evidence from a large number of cognitive and neurophysiological studies suggesting an automatic but highly context-sensitive mapping from observed on to-be-executed actions as underlying mechanism (Sebanz et al., 2006). Our low-level sensorimotor approach is in contrast with most HRI research that employ symbolic manipulation and high-level planning techniques (e.g., Breazeal et al., 2004; Alami et al., 2005; Spexard et al., 2007; Gast et al., 2009). Although it is certainly possible to encode the rules for the team performance in a logic-based framework, the logical manipulations will reduce the effectiveness that a direct decoding of others' goals and intentions through sensorimotor knowledge offers. At first glance, the motor resonance mechanism for nonverbal communication seems to be incompatible with the classical view of language as an intentional exchange of symbolic, amodal information between sender and receiver. However, assuming that like the gestural description of another person's action also a verbal description of that action has direct access to the same sensorimotor circuits allows one to bridge the two domains. In the robot ARoS, a verbal command like *Give me the short slat* first activates the representation of a corresponding motor act in ASL (e.g., pointing towards that slat) and subsequently the representation of a complementary behavior in AEL (e.g., transferring the short slat). We have introduced this direct language–action link into the control architecture not only to ground the understanding of simple commands or actions in sensorimotor experience but also to allow the robot to transmit information about its cognitive skills to the user. Verbally communicating the results of its internal action simulation and monitoring processes greatly facilitates the interaction with naive users since it helps a human to quickly adjust his/her expectations about the capacities the robot might have (Fong et al., 2003).

Our approach to more natural HRI differs not only on the level of the control architecture from more traditional approaches but also on the level of the theoretical framework used. Compared with for instance probabilistic models of cognition that have been employed in the past in similar joint construction tasks (Cuijpers et al., 2006; Hoffman and Breazeal, 2007), a dynamic approach to cognition (Schöner, 2008)

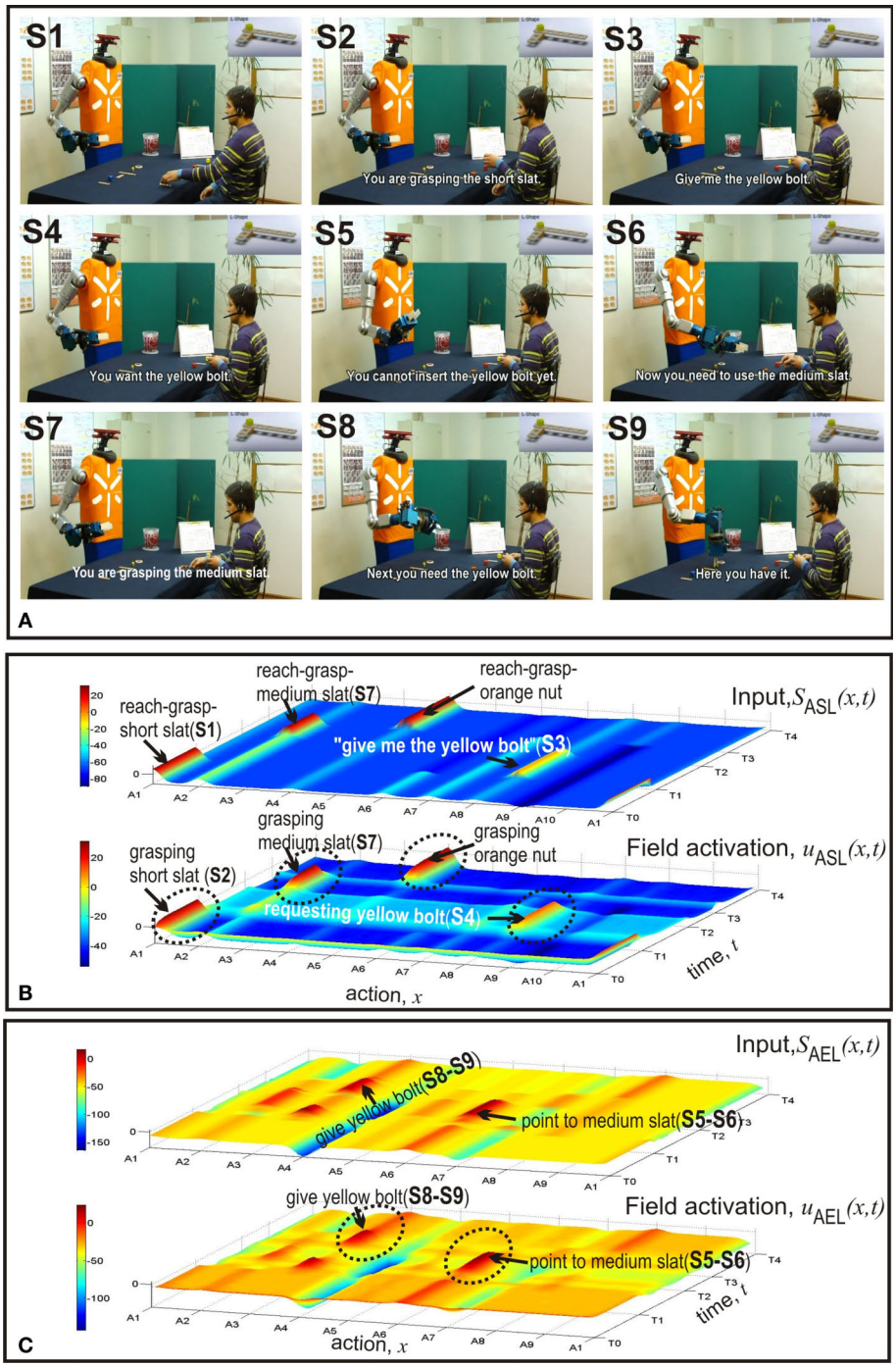

FIGURE 7 | Third example: Error detection and correction. (A) Video snapshots. **(B)** Temporal evolutions of input to ASL (top) and activity in ASL (bottom). **(C)** Temporal evolutions of input to AEL (top) and activity in AEL (bottom).

represented by the dynamic field framework allows one to directly address the important temporal aspects of action coordination (Sebanz et al., 2006). As all activity patterns in the interconnected network of neural populations evolve continuously in time with a proper time scale, a change in the time course of population activity in any layer may cause a change in the robot's behavior. For instance, converging input from vision and speech will speed up decision processes in ASL and AEL compared to the situation when only one input signal is available. Conflicting signals to ASL on the other hand will slow down the processing due to intra-field competition (compare **Figures 4 and 5**). This in turn opens a time window in which input from the AML may override a prepotent complementary behavior (**Figure 7**). We are currently exploring adaptation mechanisms of model parameters that will allow the robot to adapt to the preferences of different users.

A simple change in input strength from STKL to AEL will affect for instance whether the robot will wait for the user's explicit commands or will act in anticipation of the user's needs.

Learning and adaptation has not been a topic of the present study for which all inter-field connections were hand-coded. It is important to stress, however, that the DNF-approach is highly compatible with a Hebbian perspective on how social cognition may evolve (Keysers and Perrett, 2004). In our previous work we have applied a competitive, correlation-based learning rule to explain for instance how intention-related action chains may evolve during learning and practice (Erlhagen et al., 2006a, 2007). The interaction of the field and learning dynamics causes the emergence of new grasping populations that are linked to specific perceptual outcomes (e.g., grasping for handing over versus grasping for placing, compare Fogassi et al., 2005). Evidence from learning studies also support the plausibility of the direct action–language link implemented in the control architecture. Several groups have applied and tested in robots different neural network models to explain the evolution of neural representations that serve the dual role of processing action-related linguistic phrases and controlling the executing of these actions (Billard, 2002; Cangelosi, 2004; Wermter et al., 2004; Sugita and Tani, 2005). The results show that not only simple word–action pairs may evolve but also simple forms of syntax. A promising learning technique seems to be a covert or overt imitation of a teacher who is simultaneously providing the linguistic description. The tight coupling between learner and teacher helps to reduce the temporal uncertainty of the associations (Billard, 2002). The role of brain mechanisms that have been originally evolved for sensorimotor integration in the development of a human language faculty remains to a large extent unexplored (Arbib, 2005). We believe that combining concepts from dynamical systems theory and the idea of embodied communication constitutes a very promising line of research towards more natural and efficient HRI.

ACKNOWLEDGMENTS

The present research was conducted in the context of the fp6-IST2 EU-IP Project JAST (proj. nr. 003747) and partly financed by the FCT grants POCI/V.5/A0119/2005 and CONC-REEQ/17/2001. We would like to thank Rui Silva, Eliana Costa e Silva, Nzoji Hipolito, Toni Machado, Flora Ferreira and Emanuel Sousa for their help during the robotic experiments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/neuroscience/neurorobotics/paper/10.3389/fnbot.2010.00005/

REFERENCES

Alami, R., Clodic, A., Montreuil, V., Sisbot, E. A., and Chatila, R. (2005). "Task planning for human–robot interaction," in *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence. ACM International Conference Proceeding Series*, Vol. 121, Grenoble, 81–85.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibitory type neural fields. *Biol. Cybern.* 27, 77–87.

Arbib, M. A. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behav. Brain Sci.* 28, 105–168.

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., and Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Curr. Biol.* 16, 1818–1823.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., and Wilson, C. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends Cogn. Sci. (Regul. Ed.)* 7, 84–91.

Bicho, E., Louro, L., Hipolito, N., and Erlhagen, W. (2008). "A dynamic neural field architecture for flexible and fluent human–robot interaction," in *Proceedings of the 2008 International Conference on Cognitive Systems* (Germany: University of Karlsruhe), 179–185.

Bicho, E., Louro, L., Hipolito, N., and Erlhagen, W. (2009). "A dynamic field approach to goal inference and error monitoring for human–robot interaction," in *Proceedings of the 2009 International Symposium on New Frontiers in Human–Robot Interaction*, ed. K. Dautenhahn (Edinburgh: AISB 2009 Convention, Heriot-Watt University), 31–37.

Billard, A. (2002). "Imitation: a means to enhance learning of a synthetic proto-language in autonomous robots," in *Imitation in Animals and Artifacts*, eds K. Dautenhahn and C. L. Nehaniv (Cambridge, MA: MIT Press), 281–311.

Breazeal, C., Gray, J., and Berlin, M. (2009). An embodied cognition approach to mindreading skills for socially intelligent robots. *Int. J. Rob. Res.* 28, 656–680.

Breazeal, C., Hoffman, G., and Lockerd, A. (2004). "Teaching and working with robots as a collaboration," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*. New York, USA, 1030–1037.

Calinon, S., and Billard, A. (2008). "A framework integrating statistical and social cues to teach a humanoid robot new skills," in *Proceedings of the ICRA 2008 Workshop on Social Interaction with Intelligent Indoor Robots*, Pasadena, CA, USA, 27–34.

Cangelosi, A. (2004). "The sensorimotor basis of linguistic structure: experiments with grounded adaptive agents," in *From Animals to Animats 8*, eds S. Schaal, A. J. Ijspeert, A. Billard, S. Vijayakumar, J. Hallam, and J.-A. Meyer (Cambridge, MA: MIT Press), 487–496.

Chersi, F., Fogassi, L., Bonini, L., Erlhagen, W., Bicho, E., and Rizzolatti, G. (2007). "Modeling intentional neural chains in parietal and premotor cortex," in *Proceedings of the 37th Annual Meeting of the Society for Neuroscience, Soc. Neuroscience Abs. 636.6* San Diego.

Cohen, P., and Levesque, H. J. (1990). Intention is choice with commitment. *Artif. Intell.* 42, 213–261.

Cuijpers, R. H., van Schie, H. T., Koppen, M., Erlhagen, W., and Bekkering, H. (2006). Goals and means in action observation: a computational approach. *Neural Netw.* 19, 311–322.

di Pellegrino, G., Fadiga, I., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Exp. Brain Res.* 91, 176–180.

Erlhagen, W., and Bicho, E. (2006). The dynamic neural field approach to cognitive robotics. *J. Neural Eng.* 3, R36–R54.

Erlhagen, W., Mukovskiy, A., and Bicho, E. (2006a). A dynamic model for action understanding and goal-directed imitation. *Brain Res.* 1083, 174–188.

Erlhagen, W., Mukovskiy, A., Bicho, E., Panin, G., Kiss, C., Knoll, A., van Schie, H., and Bekkering, H. (2006b). Goal-directed imitation for robots: a bio-inspired approach to action understanding and skill learning. *Rob. Auton. Syst.* 54, 353–360.

Erlhagen, W., Mukovskiy, A., Chersi, F., and Bicho, E. (2007). "On the development of intention understanding for joint action tasks," in *6th IEEE International Conference on Development and Learning* (London: Imperial College), 140–145.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Rob. Auton. Syst.* 42, 143–166.

Foster, M. E., Giuliani, M., Mueller, T., Rickert, M., Knoll, A., Erlhagen, W., Bicho, E., Hipolito, N., and Louro, L. (2008). "Combining goal inference and natural language dialog for human–robot joint action," in *Proceedings of the 1st International Workshop on Combinations of Intelligent Methods and Applications, CIAM 2008*, Patras, Greece, 25–30.

Gast, J., Bannat, A., Rehrl, T., Wallhoff, F., Rigoll, G., Wendt, C., Schmidt, S., Popp, M., and Farber, B. (2009). "Real-time framework for multimodal human–robot interaction," in *Proceedings of the 2nd Conference on Human System Interactions 2009, HSI '09. IEEE Computer Society*, Catanid, Italy, 276–283.

Glenbach, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565.

Hamilton, A. F., and Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cereb. Cortex* 18, 1160–1168.

Hassin, R. R., Aarts, H., and Ferguson, M. J. (2005). Automatic goal inferences. *J. Exp. Soc. Psychol.* 41, 129–140.

Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron* 41, 301–307.

Hoffman, G., and Breazeal, C. (2007). Cost-based anticipatory action selection for human–robot fluency. *IEEE Trans. Robot.* 23, 952–961.

Keysers, C., and Perrett, D. (2004). Demystifying social cognition: a Hebbian perspective. *Trends Cogn. Sci. (Regul. Ed.)* 8, 501–507.

Koenig, N., Chernova, S., Jones, C., Loper, M., and Jenkins, O. C. (2008). "Hands-free interaction for human–robot teams," in *Proceedings of the ICRA 2008 Workshop on Social Interaction with Intelligent Indoor Robots*, Pasadena, CA, USA, 35–41.

McGuire, P., Fritsch, J., Ritter, H., Steil, J. J., Röthling, F., Fink, G. A., Wachsmuth, S., and Sagerer, G. (2002). "Multi-modal human-machine communication for instructing robot grasping tasks," in *2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2002. IEEE Computer Society*, Lausanne, Switzerland, 1082–1089.

Newman-Norlund, R. D., van Schie, H. T., van Zuijlen, A. M. J., and Bekkering, H. (2007). The mirror neuron system is more active during complementary compared with imitative action. *Nat. Neurosci.* 10, 817–818.

Pardowitz, M., Knopp, S., Dillmann, R., and D, Z. R. (2007). Incremental learning of tasks from user demonstrations, past experiences and vocal comments. *IEEE Trans. Syst. Man Cybern. B Cybern.* 37, 322–332.

Rizzolatti, G., and Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci.* 21, 188–194.

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192.

Schaal, S. (2007). The new robotics: towards human-centered machines. *HFSP J* 1, 115–126.

Schöner, G. (2008). "Dynamical systems approaches to cognition," in *The Cambridge Handbook of Computational Psychology*, ed. R. Sun (Cambridge University Press, New York), 101–125.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends Cogn. Sci. (Regul. Ed.)* 10, 70–76.

Spexard, T. P., Hanheide, M., and Sagerer, G. (2007). Human-oriented interaction with an anthropomorphic robot. *IEEE Trans. Robot.* 23, 852–862.

Steil, J. J., Röthling, F., Haschke, R., and Ritter, H. (2004). Situated robot earning for multi-modal instruction and imitation of grasping. *Rob. Auton. Syst.* 47, 129–141.

Sugita, Y., and Tani, J. (2005). Learning semantic combinatoriality from interaction between linguistic and behavioral processes. *Adapt. Behav.* 13, 33–52.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., and Rizzolatti, G. (2001). I know what you are doing: a neurophysiological study. *Neuron* 31, 155–165.

van Schie, H. T., Toni, I., and Bekkering, H. (2006). Comparable mechanisms for action and language: neural systems behind intentions, goals, and means. *Cortex* 42, 495–498.

Wermter, S., Weber, C., Elshaw, M., Panchev, C., Erwin, H., and Pulvermüller, F. (2004). Towards multimodal neural robot learning. *Rob. Auton. Syst.* 47, 171–175.

Westphal, G., von der Malsburg, C., and Würtz, R. P. (2008). "Feature-driven emergence of model graphs for object recognition and categorization," in *Applied Pattern Recognition, Sudies in Computational Intelligence, Vol. 91*, eds H. Bunke, A. Kandel and M. Last (Berlin/Heidelberg: Springer Verlag), 155–199.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interactions. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 358, 593–602.

Zwann, R. A., and Taylor, L. J. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *J. Exp. Psychol. Gen.* 135, 1–11.

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.

Received: 01 December 2009; paper pending published: 29 January 2010; accepted: 27 April 2010; published online: 21 May 2010.

Citation: Bicho E, Louro L and Erlhagen W (2010) Integrating verbal and nonverbal communication in a dynamic neural field architecture for human–robot interaction. Front. Neurorobot. ***4****:5. doi: 10.3389/fnbot.2010.00005*