A Unified Model of Speech and Tool Use Early Development

Sébastien Forestier[1,2] & **Pierre-Yves Oudeyer**[2,3] (sebastien.forestier@inria.fr)
Université de Bordeaux[1], Inria Bordeaux Sud-Ouest[2], Ensta-ParisTech[3], France

Abstract

Some studies hypothesize a strong interdependence between speech and tool use development in the first two years of life. To help understand the underlying mechanisms, we present the first robotic model learning both speech and tool use from scratch. It focuses on the role of one important form of body babbling where exploration is directed towards self-generated goals in free play, combined with imitation learning of a contingent caregiver. This model does not assume capabilities for complex sequencing and combinatorial planning which are often considered necessary for tool use. Yet, we show that the mechanisms in this model allow a learner to progressively discover how to grab objects with the hand, how to use objects as tools to reach further objects, how to produce vocal sounds, and how to leverage these vocal sounds to use a caregiver as a social tool to retrieve objects. Also, the discovery that certain sounds can be used as a social tool further guides vocal learning. This model predicts that the grounded exploration of objects in a social interaction scenario should accelerate infant vocal learning of accurate sounds for these objects' names.

Keywords: tool use; speech development; free play; exploration; imitation learning; social tool use; goal babbling

Introduction

Some studies hypothesize that there might be a strong interdependence between speech and tool use development in the first two years of life (Gibson, Gibson, & Ingold, 1994). Tool use and language seems to require similar information processing capabilities allowing the production and perception of sequential combinations of increasing complexity, from reaching to spoon self-feeding and from words to stories. In addition to showing similar compositional properties, speech and tool use might share some neural correlates involving Broca's area (Higuchi, Chaminade, Imamizu, & Kawato, 2009). Those common neural correlates could have an evolutionary origin in the hominid lineage, where a selection pressure for complex tool use, language and social behaviors might have together driven the increase in brain planning capabilities (Morgan et al., 2015). In particular, the development of tool use precursors follows several overlapping phases of behaviors: 1) body babbling, where babies learn to control their body parts, 2) interacting with a single object, and 3) exploring object-object interactions (Guerin, Kruger, & Kraft, 2013). From pointing movements to the control of a rake, new representations and physical understanding are developed to allow the planning of tool use actions composed of combinations of more simple actions, e.g. grasping the rake. During the same period, infants progressively learn how to efficiently use their vocal tract, comprising many complex actuators from the larynx to the lips. At birth, they produce immature protophones like squeals, growls or quasi-vowels, and by the end of their first year they are able to produce the speech-like syllables of their native language (Oller, 2014). Those syllables then form words which become the basis of syntactic combinations essential to language expressiveness. Infants do not only explore tool use and vocalizations by themselves, driven by intrinsic motivations (Moulin-Frier, Nguyen, & Oudeyer, 2013), but also spend a great part of their time interacting with their parents and other social peers where imitation is thought to be one of the important developmental pathways (Meltzoff, 1999). For instance, infants imitate the vowels produced by an adult speaker by 6 month of age (Kuhl, 2004), and 1.5-year-olds imitate demonstrations of a rake-like tool function to retrieve an out-of-reach toy (Chen & Siegler, 2000).

In order to investigate hypotheses about the joint development of speech and tool use, we seek to build an embodied model of tool use and speech learning. Existing robotic models of tool use showed first insights into how relations between tools and other objects could be learned from grounded experimentation. In (Stoytchev, 2005), a robotic arm focused on learning rake-like tool affordances from the exploration of already implemented stereotyped arm behaviors. In (Tikhanoff, Pattacini, Natale, & Metta, 2013), the iCub robot was given its arm's forward model and inverse optimization methods which led to stereotyped grasping. A recent series of robotic models considered the learning of tool use from scratch, without any kind of pre-programmed reaching skills (Forestier & Oudeyer, 2016a, 2016b, 2016c). Those models studied the developmental progression of robotic agents between phases of behaviors with objects, and the evolution of their strategies to reach goals. They have shown interesting similarities with infant development in terms of developmental trajectories and strategy choice dynamics.

Recent computational models of vocal development make use of a simulated vocal synthesizer that the learning agent must control in order to produce vocalizations, with human sounds as targets to be imitated (Warlaumont, Westermann, Buder, & Oller, 2013; Philippsen, Reinhart, & Wrede, 2014). In (Moulin-Frier et al., 2013), the agent chooses the strategy that shows the best competence progress: either autonomously training to reach phonetic goals, or trying to imitate human sounds. They show that the intrinsic motivation for learning progress self-organizes coherent infant-like developmental sequences. Those models of language acquisition study several developmental pathways to the learning of forward and inverse models of a simulated vocal tract, from autonomous exploration to human sounds imitation. However, agents are not situated into a physical environment where vocalizations have a meaning related to objects.

Several works study joint action and language learning (Cangelosi et al., 2010), but give an advanced knowledge of the linguistic interaction protocol to the learning agent who has to associate predefined actions or objects to predefined labels and learn the semantic compositionality. Also, agents learn actions without a nested tool use property.

In this paper we describe the first model that jointly considers the early development of both tool use and speech. Such a model could allow the investigation of hypotheses about the mechanisms underlying the observed links between tool use and speech development. In a previous work, we showed that the Model Babbling learning architecture (Forestier & Oudeyer, 2016b) allows the development of tool use in a robotic setup, through several fundamental ideas. First, goal babbling is a powerful form of exploration to produce a diversity of effects by self-generating goals in a task space (Baranes & Oudeyer, 2013). Second, the possible movements of each object define a task space in which to choose goals, and the different task spaces form an object-based representation that facilitates prediction and generalization, as explained by (Chang, Ullman, Torralba, & Tenenbaum, 2016). Also, cross-learning between tasks updates all skills while exploring one in particular. A novel insight was that early development of tool use could happen without a combinatorial action planning mechanism: modular goal babbling in itself allowed the emergence of nested tool use behaviors.

Here we extend this architecture so that the agent can imitate caregiver's sounds in addition to autonomously exploring. We hypothesize that these same algorithmic ingredients allow a joint unified development of speech and tool use. Our learning agent is situated in a simulated environment where a vocal tract and a robotic arm are to be explored with the help of a caregiver. The environment is composed of three toys, one stick that can be used as a tool to move toys, and a caregiver moving around. The caregiver helps in two ways. If the agent touches a toy, the caregiver produces this toy's name, but otherwise produces a distractor word as if it was talking to another adult. If the agent produces a sound close to a toy's name, the caregiver moves this toy within agent reach.

We show that our learning architecture based on Model Babbling allows agents to learn how to 1) use the robotic arm to grab a toy or a stick, 2) use the stick as a tool to get a toy, 3) learn to produce toy names with the vocal tract, 4) use these vocal skills to get the caregiver to bring a specific toy within reach, and 5) choose the most relevant of those strategies to retrieve a toy that can be out-of-reach. Also, the grounded exploration of toys accelerates the learning of the production of accurate sounds for toy names once the caregiver is able to recognize them and react by bringing them within reach, with respect to distractor sounds without any meaning in the environment. Our model is the first to allow the study of the early development of tool use and speech in a unified framework.

Methods

Learning Environment

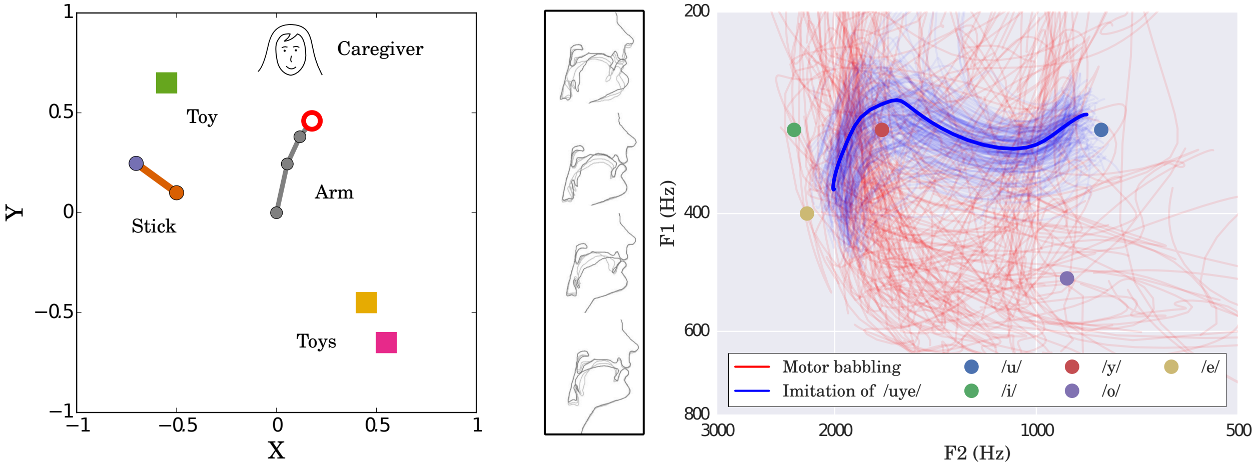
The learning environment[1] is composed of a simulated 2D robotic arm and a simulated vocal tract that the agent controls to interact with a caregiver and toys. In each trial, the agent observes the current environmental state and then executes a motor trajectory, either corresponding to moving the motors of the arm or of the vocal tract, and gets the associated sensory feedback composed of the trajectory of each object and the sound produced by the agent or the caregiver (see Fig.1).

Simulated Robotic Arm The simulated 2D robotic arm has 3 joints, with its base fixed at position $[0,0]$. Each joint rotates from $-\pi$ *rad* to π *rad* and the 3 segments of the arm have length 0.25, 0.15 and 0.1, so the arm has length 0.5. The framework of Dynamical Movement Primitives (Ijspeert, Nakanishi, Hoffmann, Pastor, & Schaal, 2013) is used to generate smooth joint trajectories from motor parameters. Each of the 3 joints is controlled by a DMP starting at the rest position of the joint (position 0) and parameterized by 7 weights: one weight on each of 6 basis functions and one weight representing the end position of the joint trajectory. To sum up, the agent provides a set of 21 trajectory parameters which are translated through DMPs to a set of smooth 50-steps trajectories for the arm's joints which gives a smooth 2D trajectory to the robotic hand.

Tool and Toys In the environment of the robotic arm, 3 toys can be grasped with the hand or with the help of a stick. The stick has length 0.25 and is considered grasped as soon as the hand reaches the handle side (orange) within a distance of 0.1. At the end of the movement the stick is dropped and stays at its current position while the arm is reset to its rest position for the next iteration. The toys are reset to a random location every 20 iterations, at a distance between 0 and 1 from the center so possibly at an unreachable position.

Simulated Vocal tract A vocal tract is simulated through the DIVA model (Guenther, 2006) and allows the production of different sounds that we can classify into vowels. In the DIVA model, a set of parameters defines a vocal tract contour where each represents one component of a Principal Component Analysis of midsagittal MRI vocal tract profiles (see Fig.1b), from the jaw and tongue to the lips position. Here we use only the first 7 articulatory parameters, controlling most of the vocal tract shape's variability. From a vocal tract contour defined by a set of parameters, the DIVA software computes the corresponding sound and outputs its first 2 formants, which are often considered to give enough information to distinguish common English vowels. The DMP framework generates smooth trajectories of vocal parameters, as described above for arm parameters, to allow the simulated vocal tract to produce simple words composed of several vowels. Each of the 7 articulators is controlled by a DMP parameterized by 4 weights: the starting and end position of the parameter trajectory, and weights on 2 basis functions. Given a set of 28 trajectory parameters provided by a learning agent, the DMPs output a set of smooth 50-steps trajectories for the 7 articulators that we use in the DIVA model, which through the DIVA software generates a smooth trajectory of the first two formants (called $F1$ and $F2$).

[1] Source code and notebooks available as a Github repository at `https://github.com/sebastien-forestier/CogSci2017`



(a) Play Environment

(b) Examples of DIVA articulators, and sound trajectories

Figure 1: Agent's robotic and vocal environment. (a) Agent's 3 DOF arm, controlled with 21 parameters, grabs toys with its hand, or uses the stick to reach toys. Caregiver brings a toy within reach if the agent says its name. (b) Agent's vocal environment representing sounds as trajectories in the two first formants space. Agent's simulated vocal tract produces sounds given 28 parameters. When agent touches a toy, caregiver says toy's name. Some sounds corresponding to random parameters are plotted in red, and some sounds produced when imitating caregiver's /uye/ word in blue (best imitation in bold, error 0.3).

Sounds: from Vowels to Words The simulated vocal tract controlled through DMPs has the potential to produce words composed of a sequence of 3 vowels in the set {/o/, /u/, /i/, /e/, /y/}. See Fig. 1 (b), "Motor babbling" condition, for an example of 200 trajectories corresponding to random sets of 28 parameters. We define a set of 6 words that the caregiver produces perfectly: {/yeo/, /euy/, /iuo/, /uye/, /eou/, /oey/}. A sound trajectory produced by the vocal tract is recognized if its distance to the perfect word is lower than 0.4.

Caregiver's guidance A simulated caregiver is given two roles to help the learning agent. First, at the beginning of the experiment, the caregiver chooses randomly a label for each toy from the set of 6 words. When the agent touches a particular toy with its hand, the caregiver then produces the sound trajectory corresponding to the label of this toy. If the agent does not touch any toy with the arm, the caregiver produces one of the distractor sounds, as if she was talking to another adult. Second, if the agent produces a sound trajectory recognized by the caregiver as the label of a toy, the caregiver moves the corresponding toy in between herself and the agent so that it becomes reachable by the agent with the hand. The caregiver is reset to a random position at each iteration.

Sensory Feedback Before choosing a motor command, the agent receives the state of the environment (or context) as the 2D position of the caregiver, the stick and the 3 toys (so 10D). At the end of the movement, the agent receives a sensory feedback s in the sensory space S (60D), from the different objects in the environment. First, the trajectory of the hand is represented as its x and y positions at 5 time points: steps 1, 13, 25, 38, 50 of the 50-steps trajectory (S_{Hand}, 10D). Similarly, the trajectories of the stick and the 3 toys during the movement are represented in 10 dimensional sensory spaces (S_{Stick}, S_{Toy_1}, S_{Toy_2}, S_{Toy_3}, 10D each). Sound, either produced by the agent or by the caregiver, is represented by the position of the first two formants at 5 time points (S_{Sound}, 10D).

Unified Modular Learning Architecture

The goal of a learning agent is to use its robotic arm and vocal tract to discover a diversity of sensory effects, and collect data to learn repertoires of skills in the form of inverse models allowing to reproduce these effects. Consequently, the agent is not given a priori a single target task to be solved, but a modular object-based representation of task spaces. The agent learns a set of sensorimotor models mapping a motor space to one particular sensory space (see Fig. 2). For instance, model 1 learns to move the hand from arm parameters, model 2 learns to move the stick, model 3, 4, and 5 learn to move one of the toys, and model 6 how to produce sounds with the arm, which will be possible by touching one of the toys with the hand so that the caregiver produces the corresponding label. Controlling vocal tract, model 7, 8 and 9 learn to move one of the toys by involving caregiver's help, and model 10 learns to produce diverse sounds autonomously.

Exploration through Model Babbling In order to actively explore and learn the 10 sensorimotor models from experimentation with the environment, learning agents use the Model Babbling architecture developed in (Forestier & Oudeyer, 2016b) that we extend to handle the 2 motor spaces: the robotic arm and the vocal tract. First, the agent performs some random exploration of motor spaces, 500 with the robotic arm and 500 with the vocal tract, to get an initial sampling of those spaces. Then, at each iteration, the learning agent first chooses to train one of the 10 models, chosen randomly (e.g. from the robotic arm to the hand sensory space). A particular goal is then randomly chosen in

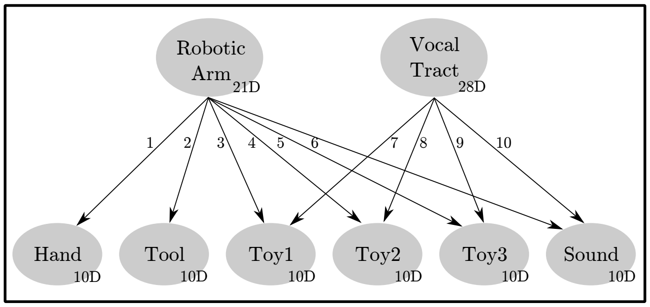


Figure 2: Learning Architecture. Agent controls 2 motor spaces and receives sensory feedback about 6 objects. Each arrow represents one of the 10 sensorimotor models learned.

the sensory space corresponding to the chosen model (e.g. a particular 2D trajectory of the hand). The agent then uses the corresponding inverse model to infer a motor command in the corresponding motor space (e.g. arm parameters) to reach the goal. Exploration happens in goal choice and in the new motor parameters that inverse models infer with generalization mechanisms and adding exploration noise.

Imitation of Sounds When the agent is choosing to train to produce sounds with its vocal tract (model 10), instead of always choosing random goals, it does this for half of the iterations (chosen randomly), and the other iterations are focused on trying to imitate the caregiver, by randomly choosing one of the sounds previously produced by the caregiver as a goal.

Forward and Inverse Models Each sensorimotor model provides a forward model and an inverse model, with the same implementation as in (Forestier & Oudeyer, 2016b). The forward model predicts which sensory trajectory would be observed given the current context and a motor command to execute. The inverse model infers a motor command that could reach a desired goal given the current context. When a motor command m is executed (either 21 parameters for the robotic arm or 28 for the vocal tract) in a context c and a sensory feedback s is received in S, all the sensorimotor models that share the same motor space are updated. New information comes as a tuple (m, c_i, s_i) with s_i being a subset of s variables corresponding to the respective sensory space, and c_i being the subset of c relevant for this sensorimotor model. The relevant context for models 1 and 10 is empty, which means that hand trajectories and vocal sounds produced by the agent do not depend on the current position of other objects. The context for model 2 is the position of the stick, and for models 3, 4, and 5, the position of the stick and of the corresponding toy. For model 6, all toys are relevant, and for models 7, 8 and 9, the caregiver and the toy is useful. Given a database of (m, c_i, s_i) experiments, an inverse model infers a probable motor command m to reach a goal s_g in a context c_i by looking for the nearest neighbor s_{NN} in S_i of s_g and retrieving the associated motor parameters m_{NN} that were used to reach s_{NN}. It then outputs m_{NN} plus Gaussian noise $(\sigma = 0.05)$ to explore new parameters.

Results

We ran 500 independent trials of 80000 iterations (or robot experiment) each. We measured how agents learned to move objects by giving them new goals in new contexts, and we analyzed the accuracy of the learned vocalizations.

Competence to Reach Toys

After 80000 iterations of training, we measured the performance of each agent to retrieve a toy depending on its current position with its preferred method: with the hand, with the stick used as a tool or involving caregiver's help. Fig. 3 shows the mean competence of all agents to retrieve toys depending on the current position of the toys. The competence error to retrieve a toy is measured by the distance between a goal trajectory given to the agent, where the toy is moved towards the center, and the actual trajectory that the agent succeeds to give to the toy. The agent chooses the strategy expected by its inverse models to best reach the goal trajectory for the toy given the current context (position of the stick, toys and caregiver) and its past experience of 80000 iterations.

In most toy locations, the normalized competence of learning agents is significantly better (46% on average) than the normalized competence of a random agent (0%). Our learning architecture thus allows to successfully reach new goals in multiple sensory spaces with multiple available strategies. Local variations reflects differences in strategy preferences and performances. For instance, where the hand cannot reach for the toy anymore, the agent still thinks this is a good strategy as it worked in a similar context (before the limit), but the hand strategy leads there to a bad performance. More training would refine the inverse models and the choice of strategy.

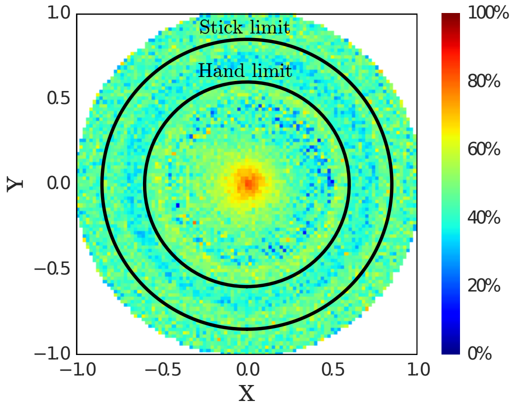


Figure 3: Competence after 80000 iterations. 0% means that competence to retrieve a toy there is as bad as with random agents, 100% says that agents perfectly retrieve a toy there.

Three Strategies to Reach Toys

Fig. 4 shows the preference for the hand, tool and vocal strategies to retrieve a toy depending on the distance of the toy. In the center region, where agents can retrieve toys with all three strategies, agents choose most often the hand strategy (around 65% of the trials) and less the other two (around 15% to 20% each). In the second region, unreachable with the hand, this strategy is still used around 50% of the trials, and the two other between 20% and 30%. In the last region where the only useful strategy is to say the name of the toy so that the caregiver brings it closer, the vocal strategy is used more often: at distance 1 from center, it is used in 49% of trials, hand strategy in 38%, and tool strategy in 13%.

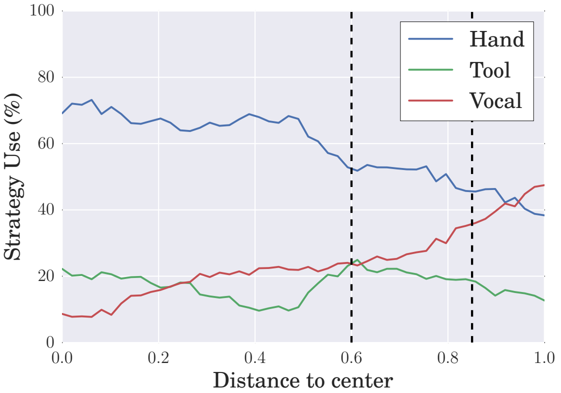


Figure 4: Strategy preferences depending on the distance of the toy. The two vertical bars shows the hand and stick limits.

Vocal Learning with Caregiver's Feedback

The agents learn to produce vocalizations both with goal babbling and imitation of the caregivers' sounds. For each agent, three of caregiver's sounds (randomly selected) are toy names and the three others are distractors: sounds that have no special meaning for the agent. We measure errors to reproduce caregiver's sounds as the distance between the sound trajectory produced by the caregiver and the best imitation of the agent. We group the results into two categories: errors of sounds that serve as toy names and as distractors. From the 500 runs we could retrieve error data for 1482 toy names and 1482 distractors. Fig. 5 shows the distribution of errors after 80000 iterations. First, 88% of sounds have an error lower than 0.4, and thus are successful imitations. Second, the median error for toy names is 0.23 and for distractors is 0.30. Imitations of toy names are more accurate than of distractors: a Mann-Whitney U test gives $p < 10^{-72}$. Errors distribution above 0.4 is similar for the two categories, but few toy name sounds have an error just below 0.4 compared to distractors: their distribution is shifted towards smaller errors.

Discussion

This unified robotic model allows to study the interaction between the early development of tool use and speech. Results show that agents learn to 1) use the robotic arm to grab a toy or a stick, 2) use the stick as a tool to get a toy, 3) learn to produce toy names with the vocal tract, 4) use these vocal skills to get the caregiver to bring a specific toy within reach, and 5) choose the most relevant of those strategies to retrieve a toy, for instance preferring to use caregiver's help when the toy is out-of-reach. Interestingly, learning the production of accurate sounds for toy names was faster than for distractor sounds because inverse models often select the use of vocalizations to retrieve toys through the caregiver. Grounding vocal interaction between agent and caregiver in a play scenario thus accelerated the learning of toys' names production.

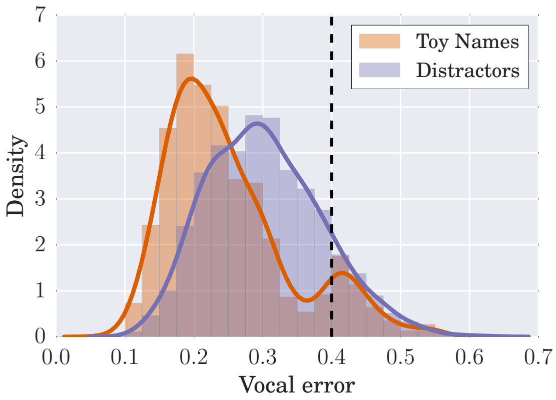


Figure 5: Distribution of accuracy of imitations of caregivers' sounds after 80000 iterations. Below 0.4 vocal error, sounds are recognized as imitations by the caregiver. Imitations of toy names are more accurate than imitations of distractors.

The proposed unified Model Babbling architecture does not integrate sequencing and combinatorial planning mechanisms and agents were not given initial teleological understanding of speech or tool use. However, with goal babbling and an object-based representation of task spaces, our architecture still allowed agents to learn behaviors showing a nested tool use structure, e.g. reusing movements of the stick to move a toy, or sound trajectories produced with the vocal tract so that the caregiver brings a toy. This suggests that observing infants using tools or asking for help with toys should not necessarily be interpreted as a correlate of capabilities for combinatorial sequencing and planning of actions.

It should be noted that for the agents in our model, involving the caregiver to move toys through vocalizations is a strategy with no special status with respect to the other strategies. This social interaction emerges from the same drive to refine sensorimotor models as in the learning of hand or stick movements. The production of sounds that can be understood by the caregiver as toy names to make it react and help can thus be interpreted as an emergent form of social tool use.

Those results offer a new prediction: exploration and play with objects in a grounded interaction scenario with a caregiver accelerates infant vocal learning of accurate sounds for the names associated to these objects. This hypothesis is consistent with experimental data from infant development research. First, (Clerkin, Hart, Rehg, Yu, & Smith, 2017) shows that the objects that are frequent in the visual field of 8 1/2 to 10 1/2 mouth-old infants are also the objects for which infants acquire the name early. They explain that the particular distribution of object frequency in visual field can help linking the heard label to the good object in a scenario where the caregiver says the name of an object. However, this data is also consistent with our hypothesis: the most frequent objects in the visual field are the ones that the infant will most often choose goals for, and will engage caregiver's help by trying to vocalize those toys' names. Infants could thus receive more vocal feedback for those words and learn to produce them earlier. This view also fits with recent data about the body-object interaction measure. In (Thill & Twomey, 2016), the authors use a measure of the extent to which adults could easily interact with a named item and show that it predicts better the age of acquisition of the name of an item than its concreteness or imageability. In other words, the easier the interaction with an object is, the sooner its name will be acquired. Furthermore, caregiver's nonvocal feedback can also help vocal learning. Indeed, (Goldstein, King, & West, 2003) provides evidence that a nonvocal feedback mechanism such as reacting to infant's vocalizations by smiling, or touching the infant can shape vocal babbling in real time. In our experiment, the caregiver reacts to a toy's name by giving the toy to the agent, which guides vocal learning. Such a mechanism could also be an important pathways to infant vocal development.

Our unified robotic model of speech and tool use gives a basis for future research in modeling interactions between their early development. From this study, we derived experimental predictions that could drive new experiments with infants and allow us to test and refine the model.

References

Baranes, A., & Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, *61*(1).

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., ... others (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, *2*(3), 167–195.

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2016). A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*.

Chen, Z., & Siegler, R. (2000). Across the great divide: Bridging the gap between understanding of toddlers and older childrens thinking. *Monographs of the Society for Research in Child Development, 65, 1, 108*.

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B*, *372*(1711).

Forestier, S., & Oudeyer, P.-Y. (2016a). Curiosity-driven development of tool use precursors: a computational model. In *38th annual conference of the cognitive science society (cogsci 2016)* (pp. 1859–1864).

Forestier, S., & Oudeyer, P.-Y. (2016b). Modular active curiosity-driven discovery of tool use. In *2016 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 3965–3972).

Forestier, S., & Oudeyer, P. Y. (2016c). Overlapping waves in tool use development: A curiosity-driven computational model. In *2016 joint ieee international conference on development and learning and epigenetic robotics (icdl-epirob)* (pp. 238–245).

Gibson, K. R., Gibson, K. R., & Ingold, T. (1994). *Tools, language and cognition in human evolution*. Cambridge University Press.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences*, *100*(13), 8030–8035.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of communication disorders*, *39*(5), 350–365.

Guerin, F., Kruger, N., & Kraft, D. (2013). A survey of the ontogeny of tool use: from sensorimotor experience to planning. *IEEE Transactions on Autonomous Mental Development*, *5*(1), 18–45.

Higuchi, S., Chaminade, T., Imamizu, H., & Kawato, M. (2009). Shared neural correlates for language and tool use in broca's area. *Neuroreport*, *20*(15), 1376–1381.

Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., & Schaal, S. (2013). Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, *25*(2), 328–373.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11).

Meltzoff, A. (1999). *Born to learn: What infants learn from watching us.* In Fox, N. & Warhol, JG (Eds.), The Role of Early Experience in Infant Development, Skillman. NJ: Pediatric Institute Publications.

Morgan, T., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S., Lewis, H., ... others (2015). Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature communications*, *6*.

Moulin-Frier, C., Nguyen, S. M., & Oudeyer, P.-Y. (2013). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Frontiers in psychology*, *4*.

Oller, D. K. (2014). *The emergence of the speech capacity*. Psychology Press.

Philippsen, A. K., Reinhart, R. F., & Wrede, B. (2014). Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In *2014 joint ieee international conferences on development and learning and epigenetic robotics (icdl-epirob)*.

Stoytchev, A. (2005). Behavior-grounded representation of tool affordances. In *Proceedings of the 2005 ieee international conference on robotics and automation. icra 2005*.

Thill, S., & Twomey, K. E. (2016). What's on the inside counts: A grounded account of concept acquisition and development. *Frontiers in psychology*, *7*.

Tikhanoff, V., Pattacini, U., Natale, L., & Metta, G. (2013). Exploring affordances and tool use on the icub. In *2013 13th ieee-ras international conference on humanoid robots (humanoids)* (pp. 130–137).

Warlaumont, A. S., Westermann, G., Buder, E. H., & Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, *38*, 64–75.