From Language to Motor Gavagai: Unified Imitation Learning of Multiple Linguistic and Non-linguistic Sensorimotor Skills

Thomas Cederborg, Pierre-Yves Oudeyer

From Language to Motor Gavagai: Unified Imitation Learning of Multiple Linguistic and Non-linguistic Sensorimotor Skills

Thomas Cederborg and Pierre-Yves Oudeyer
Inria and Ensta-ParisTech, France
http://flowers.inria.fr

***Abstract*— We identify a strong structural similarity between the Gavagai problem in language acquisition and the problem of imitation learning of multiple context-dependent sensorimotor skills from human teachers. In both cases, a learner has to resolve concurrently multiple types of ambiguities while learning how to act in response to particular contexts through the observation of a teacher's demonstrations. We argue that computational models of language acquisition and models of motor skill learning by demonstration have so far only considered distinct subsets of these types of ambiguities, leading to the use of distinct families of techniques across two loosely connected research domains. We present a computational model, mixing concepts and techniques from these two domains, involving a simulated robot learner interacting with a human teacher. Proof-of-concept experiments show that: 1) it is possible to consider simultaneously a larger set of ambiguities than considered so far in either domain; 2) this allows us to model important aspects of language acquisition and motor learning within a *single* process that does not initially separate what is "linguistic" from what is "non-linguistic". Rather, the model shows that a general form of imitation learning can allow a learner to discover channels of communication used by an ambiguous teacher, thus addressing a form of abstract Gavagai problem (ambiguity about which observed behavior is "linguistic", and in that case which modality is communicative).**
Keywords: language acquisition, sensorimotor learning, imitation learning, motor Gavagai problem, discovering linguistic channels, robot learning by demonstration.

I. Introduction

A. Gavagai Problems

Imagine a learner and a teacher in a meadow, looking in the direction of a walnut tree, with a snake on its right. The teacher utters an acoustic wave sounding like "Gavagai", while raising its two arms and opening the hands. The scene, including the meadow, the tree, the snake, the acoustic wave and the gestures of the teacher, form a context. The teacher then shows to the learner a demonstration of how to act in response to this context: he takes a round stone and throws it in direction of the snake. The stone arrives ten centimeters to the left of the snake.

From this first learning episode, several ambiguities need to be resolved by the learner:

- **Ambiguity 1): Among the many details of the scene, which aspects of the context where relevant in deciding what and how to act?** (The positions of the tree, snake? In what coordinate frame? Their color or shape? The acoustic wave sounding like "Gavagai"? The final vowel of "Gavagai"? The movements of the arms? Of the hands? The combination of the acoustic wave and the arm movement? The combination of the acoustic wave with the presence of an animal? ...).
- **Ambiguity 2): Which properties of the demonstrated action define the response, and which other ones are just irrelevant details?** (Is it important to take a round stone or could it be a square stone? or another throwable object? Is it important to reproduce the exact same trajectory of the stone? If yes, "same" in which system of coordinate? relative to the learner? teacher? tree? Or is the trajectory irrelevant, but what counts is where the stone arrives? Or maybe what is important is the purpose of the throwing, which could be for e.g. to touch or frighten the snake?).
- **Ambiguity 3): Are all relevant properties of the context and of the demonstration observable by the learner?** (for e.g. maybe "Gavagai" means "look at the animal that is undulating", and the arm-hand gestures mean "throw a stone towards the thing Gavagai tells you to look at", in which case the part of the reaction is an internal attentional operation).
- **Ambiguity 4): Did the teacher show an optimal demonstration?** (e.g. Did the stone arrive on purpose ten centimeters to the left of the snake? Or did the teacher aim at the snake but missed it?).

Resolving such ambiguities in the context of language acquisition has been called the "Gavagai problem" [1]. In this article, we will call it the "language Gavagai problem". These ambiguities, in the flow of details within demonstrations happening at a short time scale, can in principle be progressively resolved through statistical inference over multiple learning episodes and over a longer time scale, where variations of the details of the context and of the demonstration allow the learner to carve a space of interpretation hypotheses. This inference process can be constrained by a priori favoring certain hypotheses over others, but still remains challenging in general. Indeed, this sort of cross-situational learning faces additional kinds of ambiguities as novel learning episodes happen.

We would like to thank Adrien Baranes for help with the graphical presentation. This research was partially funded by ERC Grant EXPLORERS 240007 and Region Aquitaine.

Imagine a second learning episode, in the same meadow. The snake has moved to the left of the tree, and is moving further away. The teacher then utters an acoustic wave sounding like "Gavobai", while raising its two arms and opening the hands, but with not exactly the same movement timing as in the first episode. He demonstrates how to act in response to this context: he takes a piece of wood, and throws it approximately in the direction of the snake and tree. The piece of wood bumps into the tree branches, and walnuts fall on the ground.

The learner can try to use both the first and second learning episodes to identify invariants and macro-structures, and progress in the resolution of the ambiguities mentioned above. Yet, identifying "invariants" implies identifying what is similar and what is not. A difficulty is that "similarity" is not an objective property of the scene, but a measure internal to the teacher that cannot be directly observed by the learner, and thus also needs to be learnt. In particular, the second learning episode raises the following additional types of ambiguities:

- **Ambiguity 5): Is the teacher trying to teach the learner the same skill?** i.e. does the teacher consider the context and demonstrated action as a slight variation of the context and response shown in the first episode (knowing this would help to identify which are the important and irrelevant details)? Or is the context and response considered to be very different by the teacher, and possibly the relevant aspects are not the same? Is the acoustic wave "Gavobai" considered to be the same as "Gavagai" by the teacher, just pronounced a bit faster? Or is there a crucial distinction? Does the difference in the timing of the hand movements matter?
- **Ambiguity 6): Is there a sub-part of the context-response combination that is important and similar to the first learning episode, and another sub-part that is important and different?** For e.g. maybe the acoustic wave "Gavagai" and "Gavabai" are considered different by the teacher and respectively mean "look at the snake" and "look at the walnut tree", while the raising of the arms and opening of the hands means "throw something towards the thing I tell you to look at", independent of the timing of hand movements.

Such a learning scenario illustrates the diversity and depth of learning ambiguities faced by a learner trying to acquire language. But are these difficulties specific to language acquisition, or do they characterize a larger class of learning problems? We argue for the latter: these difficulties characterize a general family of problems for learning multiple context-dependent sensorimotor skills from ambiguous human teachers.

Thus, in this setting we argue that linguistic skills can be conceptualized as a particular case of such sensorimotor skills to be learnt. Indeed, in the intuitive description of the scenario above, the acoustic waves and arm gestures produced by the teacher are considered by the learner just as any other part of the global context, including current properties of the snake and tree. All the types of learning ambiguities would be kept if the learner would replace the observation of acoustic waves by the observation of the position and movement direction of the snake, and replace the observation of arm-hand gestures by the observation of the maturity of walnuts in the tree. In that case, the teacher might try to demonstrate that something should be thrown at the snake when it is close and moves towards you, and that when there is no snake danger and walnuts are mature, one shall throw something in the branches to have them fall down. There would not be here what one may call "communicative acts" in the context, and the skills would not be labelled as "linguistic", but the learning ambiguities to be resolved would be essentially the same. Hence, we propose to use the term **"Motor Gavagai problem"** to denote such imitation learning scenarii with multiple kinds of interpretation ambiguities.

As we will see in the next section, many existing models of either imitation learning of motor skills or language acquisition have not considered explicitly the full diversity of these ambiguities, but only a subset of them, and relied on the use of constraints that were different for motor and language learning. Thus, computational mechanisms and settings for learning new motor skills by demonstration have been typically quite different from mechanisms and settings for language acquisition.

After discussing related work, we will then present a computational experimental setting, associated with corresponding experiments which goal is to make a concrete step towards operationally using the structural similarity between the "language Gavagai problem" and the "motor Gavagai problem". In particular, these experiments are used as proof-of-concept that it is possible to devise a unified learning setting and architecture which allows a learner to acquire multiple non-linguistic and linguistic skills through the observation of demonstrations made by an ambiguous teacher, and without knowing a priori which skills are non-linguistic (e.g. a sensorimotor policy to be triggered as a response to particular object configurations) and linguistic (e.g. a policy triggered as a response to a speech wave or a gesture), and for the linguistic skills not knowing a priori which are the communication channels used (e.g. speech or gesture).

The computational model presented in this article is an evolution of a previous architecture presented in [2], [3], which considered a model with three agents: a teacher, an interactant and a learner. Here, the learning model has been simplified and made more generic, since no interactant is needed (but yet could be included without significant change for the learner). Furthermore, the architecture in [3] was not analyzed in terms of its ability to resolve concurrently multiple kinds of ambiguities as we do here, and its instanciation in [2] assumed a priori specific properties of a linguistic channel, which is not done here.

Concretely, as is illustrated on figure 1, in the model presented in this article, a learner will observe a set of demonstrations from a teacher. In a given demonstration, the

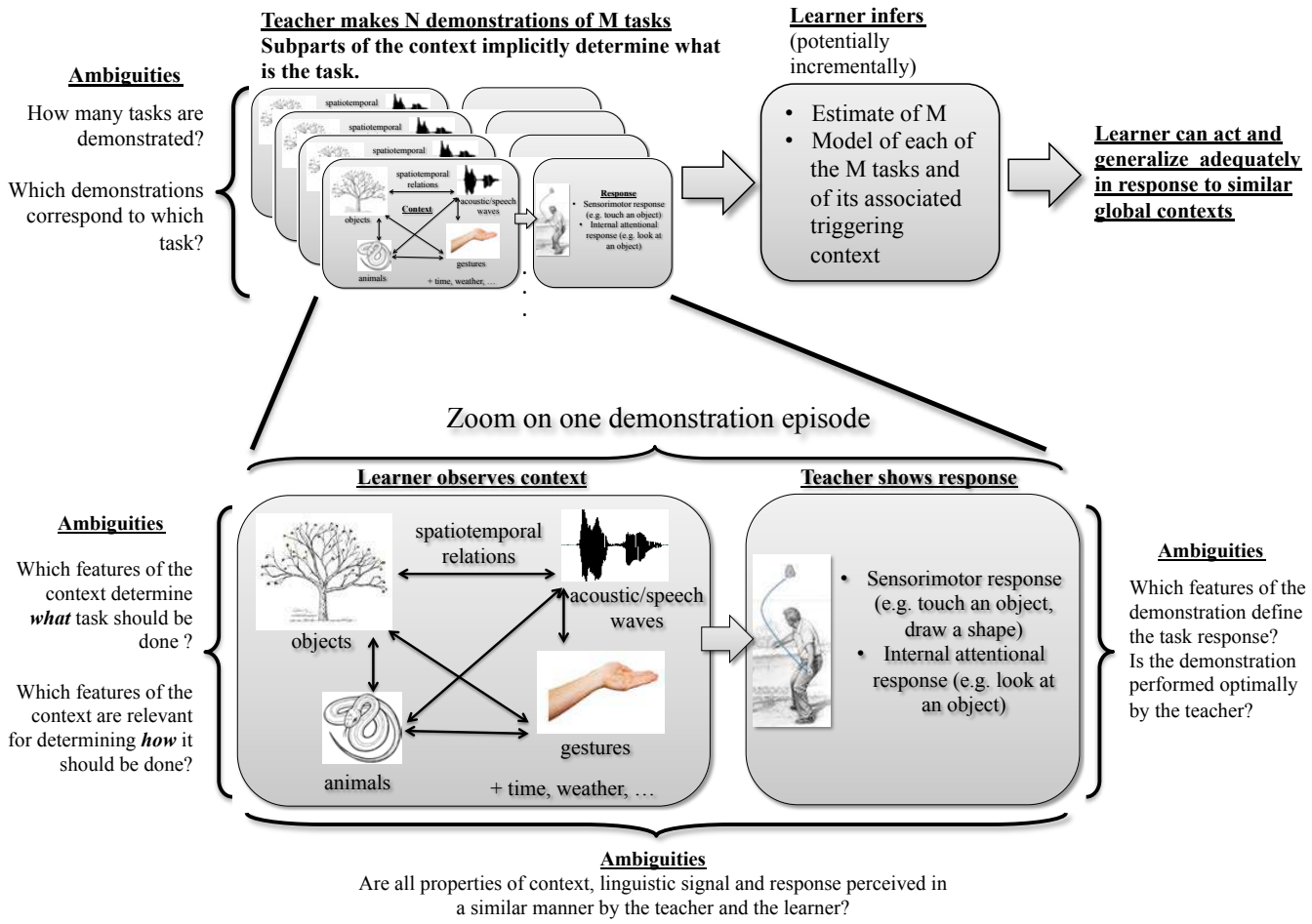

Fig. 1. **Integrated framework for learning by imitation multiple linguistic and non-linguistic context-dependant tasks**. A teacher shows a series of demonstrations to a learner. In each demonstration, the teachers sets up a global context in which the learner perceives indifferently objects, acoustic waves (which may be uttered by the teacher), gestures and their spatio-temporal relations. There is no "linguistic" signal and channel known a priori by the learner. The teacher then shows how the learner should act in response to this global context. An initially unknown subpart of the global context determines which task should be achieved as a response to this context. This subpart may in turn not be the same depending on which task-context combination is being demonstrated. It may involve speech or gestures, in which case the task can be qualified as "linguistic", or only involve object properties, in which case the task is "non-linguistic". This flexibility in learning comes at the cost that the learner has to resolve concurrently multiple kinds of ambiguities, which is shown to be feasible in the experiments presented in this article.

teacher first sets up a context, and then shows a behavior that should be executed as a response to this context (see figure 1). The context includes objects, acoustic waves and gestures (produced by the teacher), and the learner initially considers all their properties in the same way (i.e. properties of acoustic waves are not a priori different than properties of objects). Each demonstration corresponds to one of several tasks that the teacher wants to teach to the learner. Subparts of the context determine which behavior should be triggered. Some tasks consist in achieving a motor policy in a certain coordinate system, and their triggering condition depend only on object properties (speech/gesture are just considered as noise). Some other tasks consist in achieving a motor policy (possibly in a different coordinate system), and their triggering condition depend on either the details of the speech or gesture part of the context. Initially, the learner does not know how many tasks there are, and which demonstrations correspond to the same task. The learner then progressively learns how many tasks there are, and for each of them learns which part of the context determine their triggering conditions and how they should be executed, including in what coordinate system they should be encoded. In particular, the system is capable of differentiating tasks that are non-linguistic (speech and gesture are irrelevant), from linguistic tasks (which are triggered as a response to speech or gesture). It is also capable of identifying which modality corresponds to what an external observer would call a "linguistic channel" (e.g. speech or gesture).

II. Related work

The work presented in this article is related to two lines of work, which have mostly been studied independently so far. On one hand, it is related to the problem of learning context-dependent motor skills by imitation or demonstration

(in robots in particular). On the other hand, it is related to the modelling of language acquisition (in robots in particular).

A. Computational approaches to context-dependent motor learning by imitation

Computational approaches to motor learning by imitation have studied how a learner can acquire a novel context-dependent sensorimotor policy through the observation of a teacher executing this policy in response to a given context. These approaches have been especially flourishing in the area of robot learning by demonstration (see [4], [5], [6], [7], [8] for detailed reviews). The typical learning setting used is illustrated in figure 2.

The prototypical skills to be learnt in this context have typically consisted in having a robot produce coordinated movements depending on a particular (possibly dynamic) physical context such as the current body state or properties of objects in front of the robot (absolute or relative position, color, speed, etc.). In this setting the teacher wants to teach a single task to the learner. To do so, it provides a series of demonstrations. In each demonstration, the teacher sets up a context consisting of objects (e.g. positioning or throwing an object) that is perceived by the learner. Then, the teacher provides an example of a behavioral/motor policy, which is observed by the learner (e.g. a motor policy for grasping objects [9], feeding a doll with a spoon [10], performing helicopter acrobatics [11], or moving a chess piece [12]). Given a set of demonstrations of one task, the goal of the machine learner is to resolve, through statistical inference, the following ambiguities: which features of the context and demonstrated policy are relevant for defining the task response? i.e. among the many features/dimensions perceived by the learner (e.g. color, position, shape of object, speeds, successive positions, end position or effect of movement in body or external system of coordinates), which are invariant across demonstrations, and which are irrelevant details?

Various techniques have been elaborated to resolve these ambiguities and learn models of motor policies that map states (typically continuous) to actions (typically continuous). They have been ranging from regression techniques associated with dimensionality reduction algorithms (e.g. LWPR [13], GMR with PCA [9]), probabilistic approaches modelling joint distributions (e.g. [14], [15]), and neural networks trained through incremental learning or evolutionary techniques (e.g. [16]). Other techniques like inverse reinforcement learning [17], or inverse optimal control, have considered the possibility that the teacher's demonstration may not be optimal, inferring directly the "intention" of the teacher and finding an optimal policy through self-exploration [18]. While these studies have generated highly useful techniques, allowing human demonstrators to teach sophisticated motor skills to a robot, several important issues associated to the other forms of ambiguities mentioned in the previous section have so far been very little explored.

First, most studies in this area have considered learning by demonstration of a single motor task, removing ambiguities across demonstrations (i.e. is the new demonstration a variant of the same task or a new task? If it is a new task, which part of the context define its triggering conditions?). A few works have made steps towards learning of multiple tasks from unlabelled demonstrations. For example, a system combining Gaussian Mixture Regression and HMMs was shown to allow incremental learning of alternative forms of a single task, with the possibility to provide only partial demonstrations [19]. In [20], a technique based on Incremental Local Online Gaussian Mixture Regression was shown to allow for learning incrementally novel tasks by demonstration by considering various tasks as a single expandable context-dependent task. In [21], unsupervised learning techniques are used to cluster motion primitives (see also [22] for incremental motion clustering), and simultaneously make motor policy models of them.

Second, the possibility, as well as the associated challenges, to exploit multiple guiding modalities for motor learning have been so far overlooked. In particular, language can be a powerful guiding mechanisms in addition to the observation of motor demonstration for the acquisition of a novel motor skill. The work of Cakmak and Thomaz [23] for example showed how natural language dialog can allow a robot to ask questions to a human to disambiguate its interpretation of demonstrations, and Dominey et al. [24] studied how language can be used as a natural "programming" interface complementing demonstrations. Yet, in these works the meaning of words and utterances used by the teacher were all pre-programmed and known by the learner in advance, and language was explicitly used as a separate system to guide the robot. While this is highly useful for many applications, this departs from our goal to study how motor and linguistic skills can be learnt within a single process and without pre-specifying which modalities/channels are linguistic.

As we will see in the next section, some models have considered the acquisition of multiple motor tasks when linguistic labels were provided with demonstrations of each task ([25], [26], [27]), but because labels have been crisp and unambiguous, the problem attacked may be casted as several loosely coupled single task learning problems. These models have been targeting the modelling of language acquisition (rather than refined motor skills by themselves), and often make different kinds of assumptions.

B. Computational models of grounded language acquisition

Computational approaches to grounded language acquisition have considered the problem of how an embodied and situated learner can infer the meaning of utterances (forms) while observing form-meaning pairs [28], [29], [30], [31], [32], [33], [25], [26], [27], [34], [35], [2]. In many of the models in this area, form-meaning pairs are observed by the learner through interactions with a language teacher (see [33], [35], [3] for detailed reviews). These interactions can often be cast as language games [33], and provide the teacher with learning data equivalent to the process represented in figure 3: In a given context, typically defined as the configurations of the scene around the teacher and learner, the teacher produces a linguistic signal (a symbolic label, a

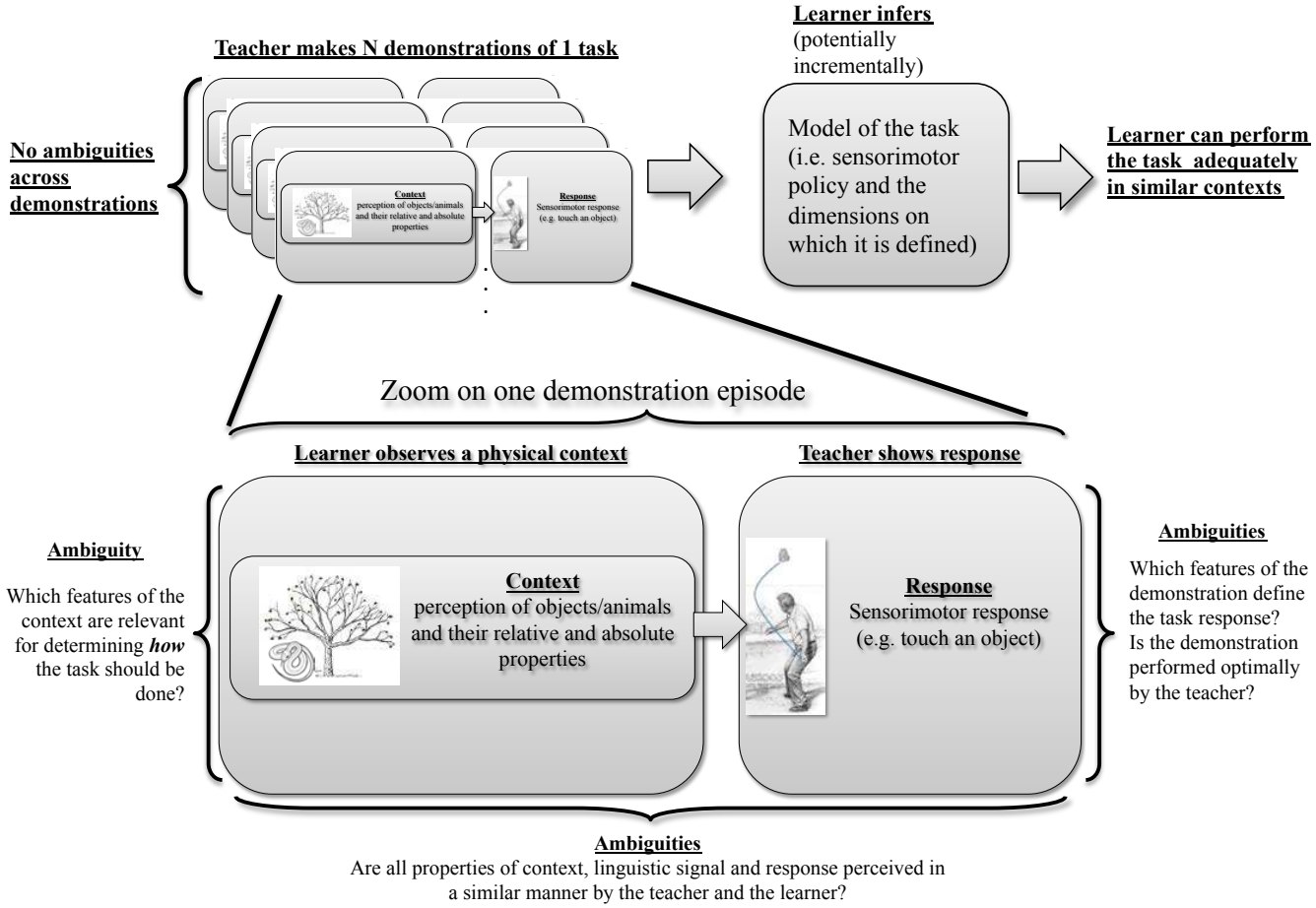

Fig. 2. **Standard architecture of computational models for imitation learning of a single context-dependent sensorimotor skill.** Here the learner knows implicitly that all demonstrations correspond to various instances of a single task. Also, the context typically only includes properties of objects in the environment: learnt skills are not linguistic. This reduces ambiguities to fasten learning, at the cost of little flexibility in learning.

speech word or a hand gesture, e.g. "square") whose meaning should be guessed by the learner, and then at the end of the language game demonstrates what the meaning of this linguistic signal is. The meaning is typically a response to the context that can be either a shift of attention towards an object referent in the scene (e.g. look at the square), or an action depending on the context (e.g. touch the square). Several such form-meaning pairs are provided to the learner, whose goal is to infer invariances across the form-meaning associations.

Several kinds of such inferences have been considered in the literature, corresponding to the resolution of various forms of ambiguities (i.e. aspects of the general Gavagai problem) as well as making various forms of assumptions. For example, some models have primarily investigated the question of how acoustic primitives in the flow of speech, i.e. phonemes, syllables and words, can be discovered with little initial phonetic knowledge and associated with simple - often crisp and discrete - meanings [36], [37], [38], [39], [40]. Some other models have assumed the existence of quasi-symbolic word representations (i.e. words are labels in the form of ascii chains, not raw acoustic waves), and focused on understanding how neural networks could learn to associate these linguistic labels with meanings expressed in terms of simple action sequences also encoded by neural networks [25], [26], [27]. Yet another family of models investigated the problem of how to guess the meaning of a new word when many hypotheses can be formed (out of a pointing gesture for example) and it is not possible to read the mind of the language teacher. Various approaches were used, such as constructivist and discriminative approaches based on social alignment [28], [29], [30], [31], [32], [33], pure statistical approaches through cross-situational learning [41], [42] or more constrained statistical approaches [34], [2]. Finally, some models have been assuming these capabilities to handle basic compositionality and have explored how more complex grammatical constructions and categories could be formed and still be grounded in sensorimotor representations [33], [41].

In spite of the richness of this landscape of models, several important issues have been little explored so far.

First, few models have attempted to consider at the same time various kinds of ambiguities. Models focusing on how to learn speech sound invariants have addressed the ambiguity "which sounds are the same and which are different" (e.g. [36], [37], [38], [39], [40]), but have considered crisp

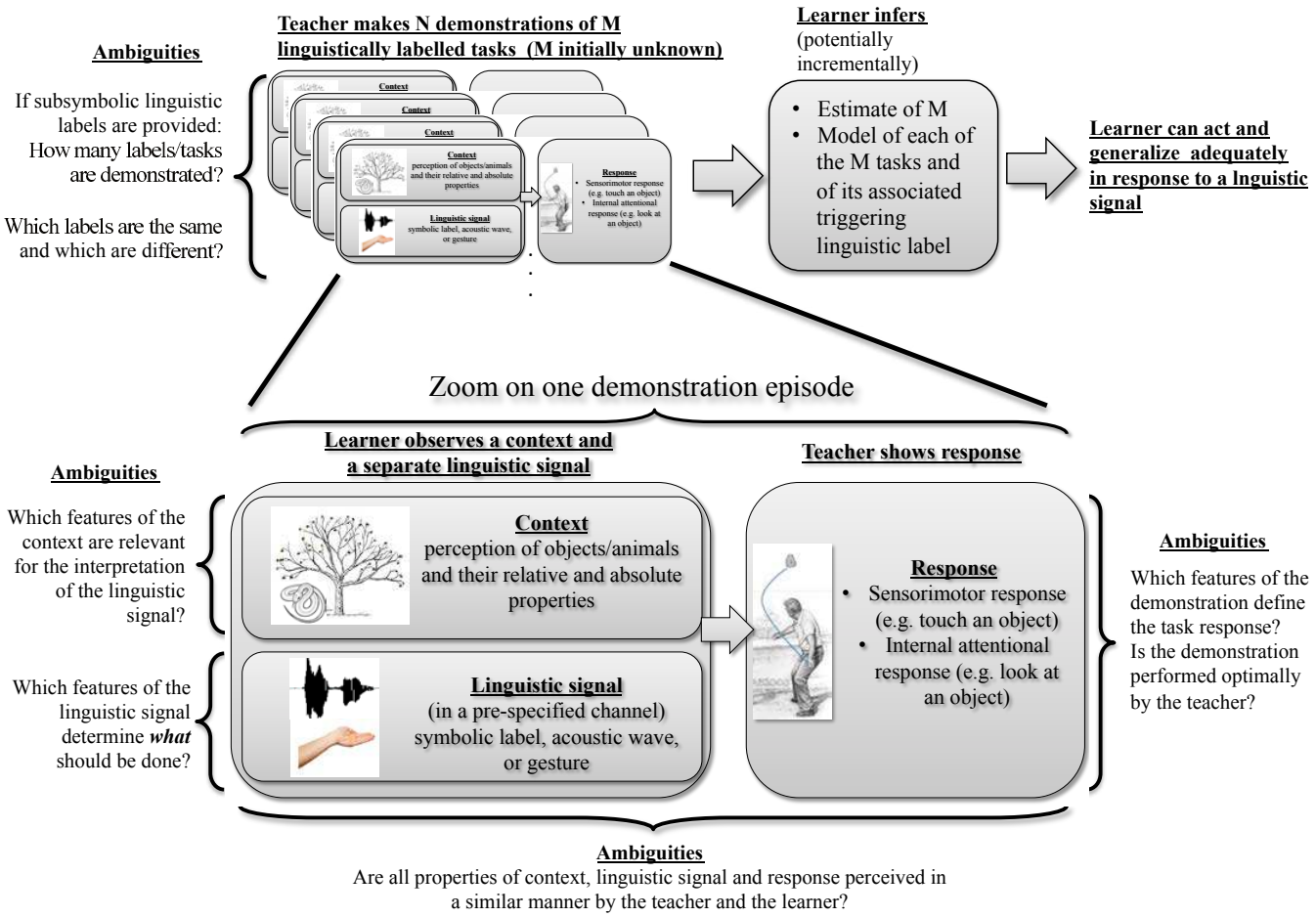

Fig. 3. **Standard architecture of computational models of grounded language acquisition.** The teacher shows form-meaning associations to the learner. The form consists of a linguistic label expressed in a pre-defined linguistic channel separate from the physical context. Meaning is a sensorimotor or cognitive response to the linguistic label combined with the context. The learner has to solve the classical language Gavagai problem, but knows implicitly that all tasks to learn are linguistic and which is the linguistic channel (e.g. speech). Such a framework does not allow the learner to acquire non-linguistic skills, or linguistic skills based on different linguistic channels (e.g. gestures or facial expression).

unambiguous meanings. Vice versa, models focusing on the inference of the meaning of words or constructions (e.g. [32], [33]) have addressed important aspects of the Gavagai problem, but still assumed that word forms were encoded as crisp unambiguous "ascii-like" labels. Still, some works considered simultaneously ambiguities in the form signal, in the meaning, and in their associations: [43] presented a multimodal learning algorithm allowing to identify speech invariants, visual scenes invariants and their associations; [44], [45] presented an unsupervised learning algorithm that allows to cluster gestures on the one hand, and actions on the other hand, and learn their co-occurrences; [46] presented an approach relying on non-negative matrix factorization to learn simultaneously speech primitives, motion primitives and their associations from the observation of a flow of speech-motion raw values. In [44], [45], another form of ambiguity is considered, which we do not address in this article: within the flow of behavior from the teacher, when do demonstrations begin and end? In another model, Lopes et al. [47] and Grizou et al. [48] considered the problem of simultaneously learning how to interpret the meaning of teaching signals provided by a human teacher (in the form of raw speech signals in [48]), and learning to solve a sequential decision problem.

Among models of language acquisition, one can also note that meanings to be learnt were mostly expressed in terms of perceptual categories (e.g. in terms of shape, color, position, etc.), and the exploration of complex action learning, where learnt action policies can be executed by the learner, has been overlooked so far. Some exceptions include [48], [44], as well as [25], [26], [27] where the focus was on how to acquire form-meaning compositional associations with a neural architecture, and little ambiguity was considered to represent simple motor policies (i.e. unlike probabilistic or regression techniques in robot learning by demonstration, the relevant dimensions defining the policy were provided to the learner). A step further was achieved in our previous work [2] where the acquisition of both compositional meanings and ambiguous motor policies (as combinatorial meanings) where considered, but prior knowledge on possible syntactic

structures to infer was assumed.

Second, some of the ambiguities described in the introductory section were in themselves not considered. To our knowledge, in all models of language acquisition so far, there is a pre-specified "linguistic channel" which is known to express the form part of a form-meaning pair to be acquired, and the learner knows what is the modality (e.g. speech) supporting this linguistic channel (see figure 3). Pre-programming this allows to avoid two kinds of ambiguities: 1) is the demonstrated task linguistic or non-linguistic (i.e. does sound or gesture produced by someone else matter for deciding what to do?); 2) in case it is linguistic, which modality is used to express a linguistic signal (speech? gestures? writing?)? Yet, understanding how to resolve such ambiguities would be highly valuable from two point of views:

- From a fundamental perspective: it would allow to study to what extent language learning can emerge as a special case of general context dependent sensorimotor learning by demonstration.
- From an application perspective: within the perspective of personal robotics, where robots will need to acquire novel skills in interaction with non-engineers, it would be very useful to be able to use a unified mechanism to teach a robot multiple tasks, some being non-linguistic, some others being linguistic, and without the need to specify for each demonstration which task it is, whether it is linguistic or not, and what is the used linguistic channel;

The learning setting and architecture we present in the next section addresses these issues, followed by the presentation of proof-of-concept experiments.

III. A Unified Architecture for Learning by Demonstration of Linguistic and Non-linguistic Skills

We present here a learning architecture which goal is to allow a learner to acquire multiple skills through the observation of ambiguous demonstrations of a teacher (summary in figures 1 and 4). This architecture is made to be a proof-of-concept of how to address concurrently the different types of ambiguities in the Gavagai problems presented in the introduction. This learning architecture integrates concepts and techniques from both previous models of language acquisition (e.g. considering the problem of learning multiple meanings/tasks in a cross-situational manner) and models of motor learning by demonstration (e.g. considering meanings as complex sensorimotor policies whose coordinate systems must be inferred). It also extends them by considering the problem of learning within a single process linguistic and non-linguistic skills, and without formally pre-specifying a "linguistic" channel.

A key idea is to have the learner consider a generalized context which includes behaviors of peers, such as speech waves and gestures, as elements initially similar to other properties of the scene such as object properties. The learner has then to infer which demonstrated skill depend on which subpart of a generalized context: some skill will depend only on object properties (they will be called "non-linguistic"; some other skills will depend on speech waves produced by a peer; some other skills will depend on gestures produced by a peer.

We now detail the learning situation as considered in the computational model.

A. Learning situation

As illustrated by figures 1 and 4, we consider a learner observing a teacher providing demonstrations of how to act in given contexts. In a given demonstration, the teacher first sets up a context, which is perceived by the learner. In experiment 1 below, this context consists of an object position, a speech wave that he produces, and a hand starting position (perceived in several systems of coordinates, here called "framings"). In experiment 2, the context additionally includes a hand gesture made before taking the starting position. Then, after the context is set, the teacher shows to the learner how to act. In both experiments below, action consists of a hand movement, i.e. a motor policy represented by the learner as a probabilistic mapping between the current state of the context and the speed of the hand.

When observing a single demonstration, the learner is faced with the following ambiguities: Which features/dimensions of the context determine what should be done (i.e. which details were important and which were not important for triggering the observed motor policy)? Which features of the context are relevant for determining how the action should be done: which feature of the demonstration define the task response?

Yet, the teacher does not provide a single demonstration, but multiple demonstrations. Multiple demonstrations allow to make cross-situational statistics, but also pose novel ambiguities that the learner has to resolve: how many underlying tasks are shown by the teacher? Which demonstrations are demonstrations of the same task? Indeed, the teacher does not provide external "labelling" information with each new demonstration: what should be done, and how it should be done is determined by sub-parts in the continuously perceived context, and these sub-parts correspond to underlying (noisy) invariants which need to be inferred by the learner. Furthermore, different tasks may have different corresponding relevant sub-parts of the context defining when to trigger them and how to achieve them.

B. Perception and sensorimotor apparatus in experiments

The two experiments presented below include an artificial learner with a simulated hand, a real human demonstrator, and a simulated object, modeling the learning situation depicted on figure 4. In the actual experiment, the robot is not physical, but simulated on a screen, and the human interacts with him through the use of a mouse (to demonstrate hand movements and to produce gestures) and a microphone. The simulated hand and object live in a 2D plane on a screen for the sake of visual illustration of the results.

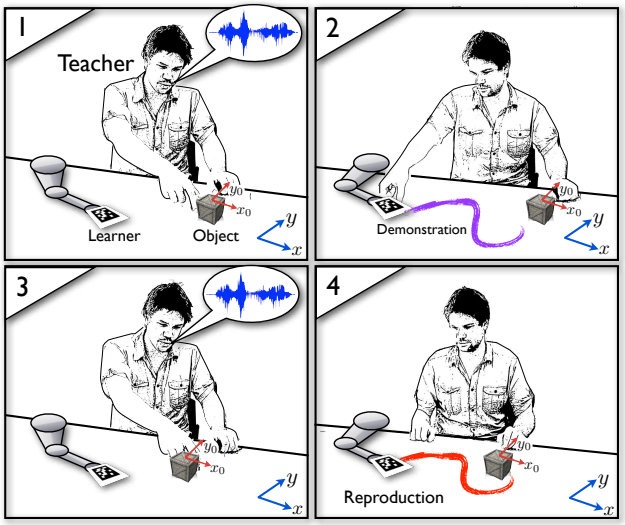

Fig. 4. The learning situation that is simulated in the experiments. **In one demonstration episode (boxes (1) and (2))**, the teacher sets up a context (1) by placing an object and producing a speech wave utterance (or a hand gesture in experiment 2). Then, the teacher performs a demonstration (2), taking the hand of the learner to show him which movement (i.e. policy for controlling the hand movement) should be produced as a response to such a context. The teacher shows many demonstration episodes, where multiple contexts (object placement, speech, gestures) and kinds of responses are demonstrated. The learner has to infer regularities to resolve the multiple kinds of ambiguities shown earlier on figure 1 (for example, the learner does not know initially if a particular speech wave or gesture matter for deciding what to do). **In the testing phase (boxes (3) and (4))**, the teacher sets up a similar context (3), placing the object and uttering a speech wave (or producing a gesture) which may not be exactly the same as during demonstration. In (4), the learner has to produce autonomously an adequate movement in response to such a context.

1) Context: In each demonstration, the teacher sets up a context perceived in the following manner by the learner:

- **Object position** $posO$**:** Absolute coordinates of the 2D position of the object (represented as a point);
- **Hand starting position** $posH_f$**:** Starting position of the hand at the beginning of the movement demonstration, provided by the teacher with the mouse. The learner perceives this starting position in 3 systems of coordinates, called framings. In experiment 1, f_1 encodes the position of the hand in an absolute frame of reference (same as O). f_2 is an object centered referential. f_3 is redundant and concatenates f_1 and f_2. In experiment 2, f_3 is replaced by a referential centered on the hand starting position.
- **Speech waves** S**:** Speech sound produced by the teacher (in experiment 1 and 2, these are instances of the words "flowers", "triangle", "point","dubleve" (french pronunciation of "w") and "circle"). This is presented to the learner as a raw speech wave, captured from a human with a standard laptop microphone, which he first transforms as a series of Mel Frequency Cepstral Coefficients vectors, and then projects onto a 3 dimensional manifold with a projector that is independent of the experiment. This projection is made using a kernel with three prototype acoustic waves (generated through a variant of K-means operating on a pre-recorded independent speech episode) and equipped with the DTW similarity measure. This allows to represent speech waves in the context as a fixed length vector homogeneous to the representation of object and hand properties. This is here chosen to be 3 dimensional for purpose of visual illustration in figures, but higher dimensional fixed length representations may be used, for example using the bag-of-features approach presented in [49], [50]. If the teacher does not produce any sound, then noise recorded in the human demonstrator environment is perceived by the learner.
- **Hand gestures** G**:** In experiment 2, the teacher also produces a gesture as part of the context. This gesture is produced as a movement of its hand captured by the mouse. The learner perceives each gesture as a sequence of 2D positions, which it then projects onto a 3 dimensional manifold with a projector that is independent of the experiment. The mathematical projector used is the same as for the speech, except that here prototypes in the kernel were taken as the results of k-means applied to a set of movements unrelated to the experiment. If the teacher does not produce a gesture, then a random point in this 3D manifold is generated. Like for the representation of speech, the choice of such a projector is motivated by visual illustration. Because it drastically reduces dimensions, it adds additional ambiguity in the signal, which may be diminished by using more sophisticated projections.

2) Action demonstrations and reproductions: When a context is set, the teacher provides a demonstration of how to act in response to this context. The response consists here in a motor policy driving movements of the hand. Starting from the hand initial positions, the teacher executes the movement using a mouse.

The movement is perceived by the learner as a sequence of hand positions (in each of the three framings) associated to the current context state $(posO, posH_f, S, G)$, which includes the current hand position. This is then transformed by the learner into a representation suited for reproduction and generalization of the sensorimotor policy: a given demonstrated movement is transformed as a series of associations between the current state $(posO, posH_f, S, G)$ and the speed of the hand $\delta posH_f$. As shown in the algorithms below, the grouping of several such series of data across demonstrations, corresponding to the same inferred task, will allow the learner to make a probabilistic model (using incremental local Gaussian Mixture Regression) of the corresponding closed-loop policy and generalize what should be done even in states of the context never exactly encountered during demonstrations:

$$(posO, posH_f, S, G) \rightarrow \delta posH_f$$

C. Algorithmic Architecture

The algorithmic architecture is divided into a learning sub-architecture doing off line analysis of the data provided by the set of demonstrations, and a reproduction

sub-architecture that uses the results of this analysis for computing online how to act in response to the current state of the context. During reproduction, the teacher builds a context (e.g. sets up the object, produces a speech sound and/or a gesture), and the learner has to produce the adequate sensorimotor policy as a response.

D. Learning algorithm

The learning algorithm takes the demonstrated hand trajectories and the generalized context as input, and creates estimates of which demonstrations are instances of the same task, and which is the correct framing for each such task, that are later used by the reproduction algorithm. We present here the outline of the algorithm. Further details of the grouping algorithm are given in appendix VI.

1) Similarity estimation: The goal of the similarity estimation step is to measure similarity of demonstrated hand motor policies across all pairs of demonstrations in the demonstration set. The similarity $\Delta_{m,n,f}$ between demonstrated motor policy of demonstration m and demonstrated policy of demonstration n computes the average difference between hand speeds over all observed states in demonstration m (thus is asymmetric): $\Delta_{mn} = \sum \delta_i^2$, where δ_i is the difference in output between point i of demonstration m and the point of demonstration n closest to its associated current context. Since the closeness of context depends on framing (because it includes $posH_f$), the similarity is also dependent on framing f. We assume here that each demonstration corresponds to a single task and a single framing.

2) Grouping algorithm: Estimation of task groups, policy framings and triggering contexts: The grouping algorithm takes the estimated similarities as inputs and outputs an estimated set of groups gathering demonstrations that have a high probability to be of the same task. For N trajectories, this is an NxN matrix P where p_{it} is the probability that trajectory number i is an instance of task group number t, since we know there is at most N different tasks given N demonstrations.

The main assumption used in this grouping algorithm is that demonstrated hand movements with high similarity in a given framing are more likely to be instances of the same task. The details of the grouping algorithm can bee seen in appendix VI-B. This grouping algorithm is a form of Expectation-Maximization algorithm suited to the problem we address. Intuitively, it searches for a grouping and an associated local measure of similarity within each group, in order to maximize the measures of similarity of demonstrations within one group and dissimilarity across group. Interestingly, because similarity measures both consider the set of states actually observed in given demonstrations, and their framing (each group gets associated with the framing maximizing the intra-group similarity), each group ends up having its own different local measure of similarity which characterize both which subpart of the context is relevant to the corresponding task, and which framing is associated to the sensorimotor policy to be executed. Thus, the result of this grouping is manifold: an estimation of task groups, of the framing to be used for motor policy representation in each task, and of the triggering contexts for each group (see reproduction below).

While the grouping algorithm is achieving well its goal within this article, we do not claim it is optimal. The grouping algorithm is currently a batch computation. It is however well suited for an incremental version (with current data it takes only a few seconds on a modern laptop but with larger number of tasks, demonstrations and number of possible framings, time could become a problem). When the algorithm has grouped all the observed demonstrations and found the corresponding framings, it can use this information when new demonstrations are added. If a new demonstration is similar to one of the established groups, when viewed in that group's preferred framing, then it can simply be added. Otherwise the membership values already found can be reused to bootstrap a new incremental optimization.

E. Reproduction algorithm

After demonstrations provided by the teacher have been processed by the learning algorithm, the learner is tested. The teacher performs a series of tests. In a test, the teacher first sets up a context (including potentially producing a speech sound or a gesture) from which the learner has to produce the appropriate sensorimotor policy. The learner operates in two steps.

a) Group selection: First, the context observed by the learner is used to decide to which group (i.e. what task) built during learning it corresponds. This is achieved by choosing the group that has the highest probability to be associated to such a context. This uses a probabilistic model of context distributions within one group, which models explicitly the relative importance of context features for determining this probability (i.e. object position? speech? gesture?). For each group t of demonstrations the mean μ_{dt} and variance σ_{dt}^2 of the data in dimension d is calculated for each dimension. p_{dt} is now the probability density of the gaussian with μ_{dt} and σ_{dt}^2 at the current context S. To determine what task is to be executed in the current context S, each task grouping t gets a relevance score $R_t = p_{1t} \times p_{2t} \times ... \times p_{Dt}$. The task with the highest relevance score R_t is selected and the data of that group (seen in the framing of that group) is used to build local models during the entire reproduction.

b) Online generation of motor commands: Once the adequate grouping has been determined, the learner builds online a probabilistic model of the mapping

$$(posO, posH_f, S, G) \rightarrow \delta posH_f$$

, using the framing associated to this group and all the points associating particular speed commands to particular states over all demonstrations of the group. This model is used online to compute what action to make at each time step (i.e. speed command to change the x and y positions of the hand), and for a pre-determined time duration. The algorithm used to build this mapping is Incremental Local Online Gaussian Mixture Regression (ILO-GMR), introduced in [20], [51]. ILO-GMR is a variation of the Gaussian Mixture Regression

method (GMR) [52] [14], which was itself shown to be highly efficient for real world high-dimensional robot learning by demonstration, allowing to detect which dimensions were important and which were not in given parts of the state space and given a set of demonstrations corresponding to a single task. ILO-GMR extends GMR by allowing fast incremental learning without using an EM algorithm to recompute a GMM. Like GMR, ILO-GMR is capable of identifying automatically the relative weight of various feature dimensions in the human movement demonstrations (thus it complements the automatic identification of framing described earlier, which removed irrelevant dimensions).

IV. Experiments

We here show the results of two series of experiments. In the first experiment, the teacher is showing unlabelled demonstrations of five tasks, some where the triggering condition depends on the speech sound he utters (linguistic tasks), and some where the triggering condition depends only on the object position (non linguistic tasks). In the second experiment, the teacher produces gestures in addition to speech sounds. He shows unlabelled demonstrations of seven tasks: some where the triggering conditions depend on the speech sound he utters (linguistic tasks with speech as the communicative modality), some where the triggering conditions depend on the gesture he produces (linguistic tasks with gesture as the communicative modality), and some where neither speech nor gesture matter to decide what to do (non linguistic tasks).

In each experiment, the performance of the learner is evaluated by (i) comparing the estimated task groupings of the learning algorithm with the actual task identities of the demonstrations (ii) comparing the estimated framings with the actual framings (iii) comparing the task group selected with the intended task during the reproduction/test phase (iv) by comparing the reproduced hand movements with the task description and the corresponding demonstrations.

A. Experiment 1: Learning Multiple Tasks from Unlabelled Demonstrations with a Context Including Speech

1) Tasks Demonstrations: In this experiment, the teacher can produce a speech sound in addition to setting up an initial object and hand position (but sometimes does not produce speech at all). The learner perceives the demonstrated hand movements (hence can encode policies) in 3 coordinate systems (framings): f_1 encodes the position of the hand in an absolute frame of reference (same as O). f_2 is an object centered referential. f_3 is redundant and concatenates f_1 and f_2. Five different tasks are taught by the teacher at the same time, corresponding to the following types of demonstrations (the teacher does not tell the learner of which type a given demonstration is):

- **Task a)** The teacher utters an instance of an acoustic wave consisting of the word "flower", and whatever the object position, shows the corresponding response: he encircles the object counter clockwise (task defined in framing 2).
- **Task b)** The teacher utters an instance of an acoustic wave consisting of the word "triangle", and whatever the object position, shows the corresponding response: he draws a triangle clockwise to the left of the robot (task defined in framing 1).
- **Task c)** The teacher utters an instance of an acoustic wave consisting of the word "point", and whatever the object position, shows the corresponding response: he draws a big square clockwise (task defined in framing 1).
- **Task d)** The teacher utters no sound, and places the object close to the robot and to the right, and shows the corresponding response: he draws a small square counter clockwise with the bottom right corner at the object (no matter what the speech input is) (task defined in framing 2).
- **Task e)** The teacher utters no sound, and places the object close to the robot and to the left, and shows the corresponding response: he encircles counter clockwise the point (0,0) in the fixed reference frame no matter what the speech input is (task defined in framing 1). The policy in this task is identical to the policy in task a) in that it is to encircle the point (0,0), with the only difference that the reference frame is different (besides different relative starting positions the demonstrations of task a in framing 2 looks just like the demonstrations of task e) in framing 1).

Four demonstrations of each task were provided and presented to the robot unlabelled. For the 3 linguistic tasks (tasks a, b and c) the object position distribution was uniformly distributed over the intervals: $-1 < x < 1, 1 < y < 2$, and for the 2 non linguistic tasks the object y positions were drawn from the uniform distribution $-1.25 < y < -0.5$ and the x positions were drawn from $-1 < x < -0.25$ for task d and $.25 < x < 2$ for task e. The starting hand position (demonstration and reproduction) is always drawn uniformly from $-0.25 < x < 0.25, -1.5 < y < -1.25$.

2) Results: In figure 5 we can see the results of the grouping algorithm in addition to the set of demonstrated trajectories. The demonstrations have been sorted into 5 groups with 4 demonstrations each. We can also see what framing was estimated for each group (marked with a * in the figure). We can see that each group contains hand trajectories that correspond to one of the tasks descriptions (which task the trajectories correspond to is indicated to the left in the figure). Each of the estimated task groups has four speech points and four object positions shown in the two columns to the right.

In figure 6 we can see each of the 20 reproductions individually, with the imitator's estimate of the currently appropriate framing seen in the top left of each reproduction. In two of the reproductions of task b), the imitator completes a few correct laps around the triangle, but then starts drifting into the middle. Otherwise we can see that the tasks are reproduced adequately if we compare the reproductions with the task descriptions and the demonstrations shown in figure 5.

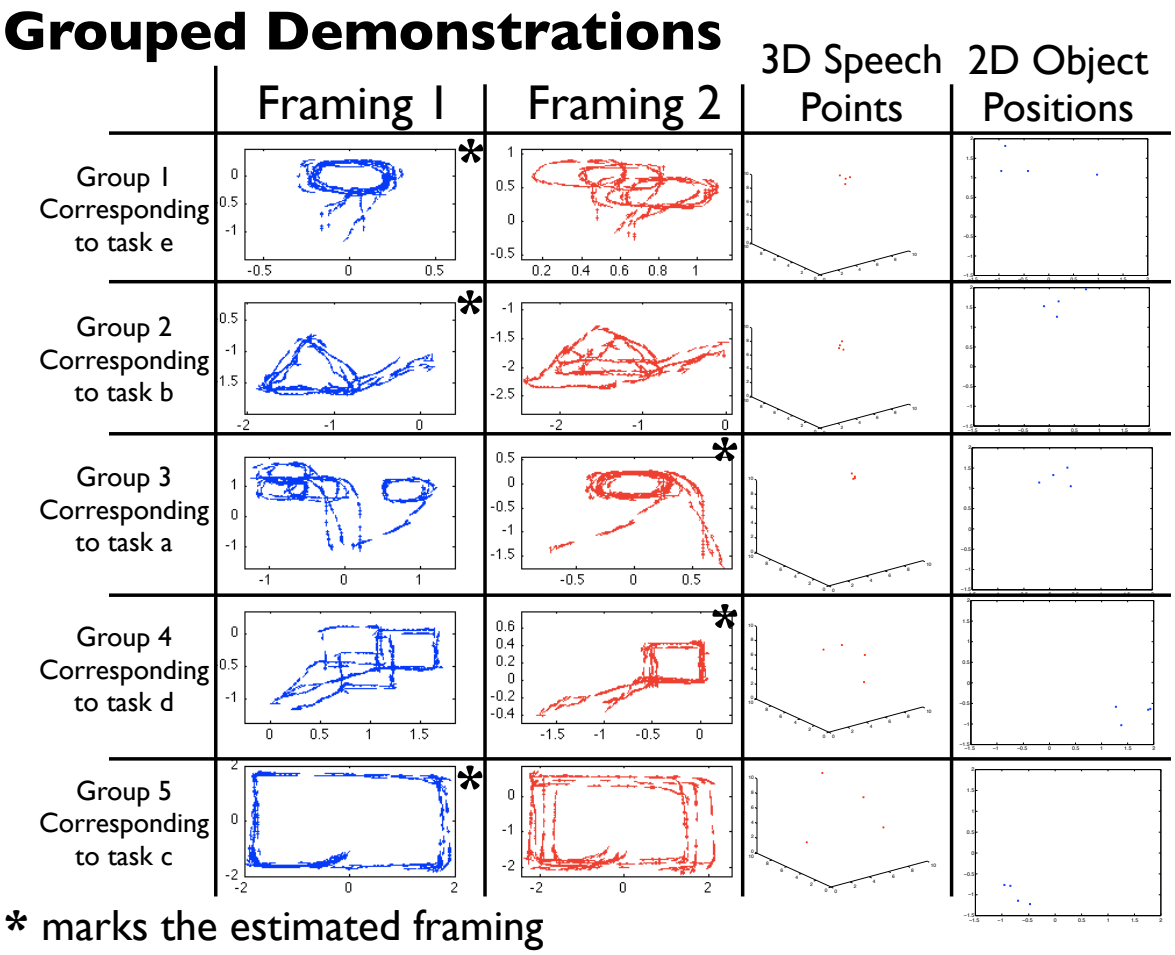

Fig. 5. This shows the result of the grouping algorithm applied to the data in experiment 1. The five task groups that were found are shown in framing 1 (absolute frame of reference) and in framing 2 (object centered). The stars indicate which frame of coordinate is inferred to be the right one for each group. In the two columns on the right, the projections of perceived speech waves onto a 3D manifold is shown (when the teacher utters no sound, the sound in the environment is still perceived), as well as the perceived 2D object positions in each demonstration episode. We can see that each group found does correspond to one of the tasks described in the task descriptions, and we can also see that the correct framings were found (also notice that the demonstrations look more coherent when viewed in the correct framing). The ordering of the tasks are random and will be different each time (this time it is e,b,a,d,c) but each time the same set of region-framing-data tuples are found. In order to avoid duplication, this figure also serves to show what was demonstrated (since each of the task groups found consists of the demonstrations of one task, the only difference of showing the task demonstrations separately would be in the ordering).

B. Experiment 2: Inferring the Linguistic Channel Among Multiple Modalities

1) Tasks Demonstrations: In this experiment, the teacher can produce either a speech sound or a gesture in addition to setting up an initial object and hand position. The learner perceives the demonstrated hand movements (hence can encode policies) in 3 coordinate systems (framings): f_r encodes the position of the hand in an absolute frame of reference. f_o is an object centered referential. f_s is a referential centered on the hand starting position. Seven different tasks are taught by the teacher at the same time. Two of the tasks are to be performed as a response to a specific object position, two tasks as a response to a speech command, two as a response to a gesture and one should be triggered when the robot hand starting position is in a certain zone. As in experiment one, it never occurs, neither during demonstration or reproductions, that the context contains two such conditions. The seven tasks are:

- **Task 1)** The teacher utters no sound, produces no gesture, and places the object to the left, and shows the corresponding response: he draws an L shape (framing f_s).
- **Task 2)** The teacher utters no sound, produces no gesture, and places the object to the right, and shows the corresponding response: he draws an R shape. Tasks 1 and 2 are meant to demonstrate that it is possible to learn to generate a gesture as a response to a world state (something that might look like a symbolic description of the world to an external observer) (framing f_s).
- **Task 3)** The teacher utters an instance of an acoustic wave consisting of the word ”dubleve” (French for w), and shows the corresponding response: he draws a W shape (framing f_s).
- **Task 4)** The teacher utters an instance of an acoustic

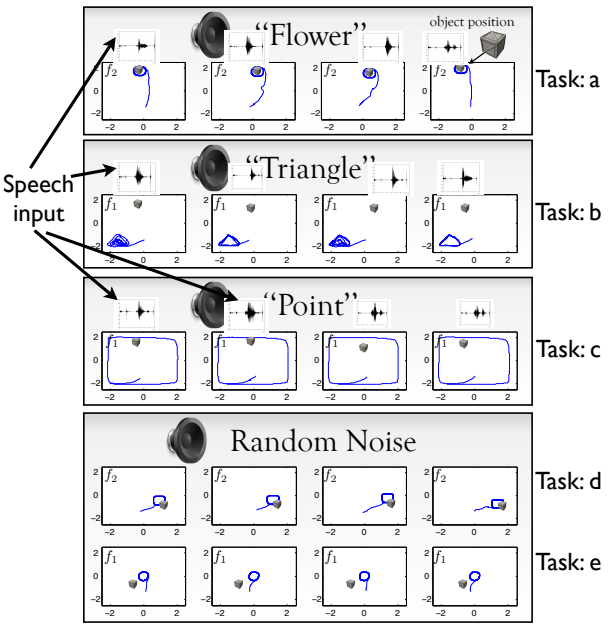

Fig. 6. **Experiment 1: test phase**. Each lines shows four examples of motor responses learnt and produced by the learner in the test phase in response to global contexts corresponding to each of the five tasks of experiment 1 (like during the demonstration phase, no category label is provided to the learner to indicate which task the context affords, this is inferred by the learner). For each task, the inferred frame of reference is also indicated (f_1 or f_2). Comparing to the demonstrations and the task descriptions, we can see that the reproduced trajectories correspond reasonably well with what the imitator was supposed to do. We see that the triangle task (task b) has a tendency to sometimes go into the middle of the triangle and circulate in a deformed trajectory after having completed a few correct laps. This problem is not due to the grouping algorithm and demonstrates a shortcoming of the ILO-GMR algorithm.

wave consisting of the word "circle", and shows the corresponding response: he goes around in a circle around the point $(0,0)$ in the reference frame of the robot. Tasks 3 and 4 shows that verbal commands can be used either to draw a shape that may be considered as a symbol by an external observer, or perform a more traditional action policy (framing f_r).

- **Task 5)** The teacher produces an instance of an "S" shaped gesture, and shows the corresponding response: he goes around in a square with the lower left corner of the square coinciding with the object (framing f_o).
- **Task 6)** The teacher produces an instance of a "P" shaped gesture, and shows the corresponding response: he pushes the object (framing f_o). Tasks 5 and 6 tasks shows that it is possible for the architecture to handle what can be seen as different forms of symbolic communication by an external observer: a gesture can also be used to command an action. In these two tasks the approximate shape of the gesture determines what to do so it might look symbolic; as long as the shape is similar to "S" the square task is performed, and the exact shape have no influence on how it is performed. The position of the objects also affects the task execution but here it smoothly modifies the policy.
- **Task 7)** When the starting position of its hand position is far away from $(0,0)$, then the teacher shows the appropriate response: go to the point $(0,0)$ (framing f_o)

The set of demonstrations provided by the teacher can be seen in figure 7.

2) Results: The results of the grouping algorithm can be seen in figure 8. In order to make viewing of the results easier the first four demonstrations (1 to 4) are of the first task, the next four demonstrations (5 to 8) are of the second task, and so on. The fact that they are demonstrated in this pattern has no impact on the algorithm but makes it possible to immediately determine visually if the algorithm was successful. The demonstrations of task 7 is not identified as a task, which is a failure of the algorithm, but the demonstrations of the other 6 tasks are grouped correctly.

Why task 7 is not grouped correctly The 4 demonstrations of task 7 can be seen in figure 9 in different colors. The green demonstration might look similar to the other demonstrations from a human observers point of view. However, to the policy similarity measure defined it is actually quite different from the red and blue demonstrations (the red demonstration is actually more similar to the demonstrations of task 1 than to the green demonstration). In these four demonstrations, the actions taken are not very similar in any of the framings hypothesized since they are reaching the same point from different directions. The framing for this task is the coordinate system relative to the robot (framing 1) and this input is indeed all that is needed to define a consistent policy. For the grouping algorithm to see the policies as similar it would however be necessary to view the output in terms of speed towards the point $posHx_r = 0$, $posHy_r = 0$, or movement in a coordinate system with one axis intersecting the starting position and the point $posHx_r = 0$, $posHy_r = 0$ as suggested in [53]. If there are intermediate demonstrations the grouping algorithm can succeed anyway according to the principle A is similar to B and C, B is similar to A,C and D, C is similar to A,B,D, and E and D is similar to B,C E and F, E is similar to C,D and F. The starting positions are generated randomly and often the demonstrations will be similar enough to be grouped together. With these 4 specific demonstrations it sometimes happens that demonstrations 1, 2 and 3 or demonstrations 1 and 2 are grouped together to form a task. The proper way to fix this problem would be to either provide a framing where the demonstrations look the same or to give the imitator the ability to find such a framing by itself.

Finding back the correct task and the correct framing from the current state. For the 6 tasks that are correctly grouped, the reproductions are successful except for around 5% failure rate for task 4. This is a different type of problem than the grouping problem of task 7, and it comes from the fact that the learner does not know which part of the context is relevant. If considering only the relevant part of the context, task 4 is always found correctly. But since the learner does not know what part is relevant, around one in 20 times, the other aspects of the context is just much more similar to what it was during the demonstrations of task 3. This type of problem decreases with more demonstrations,

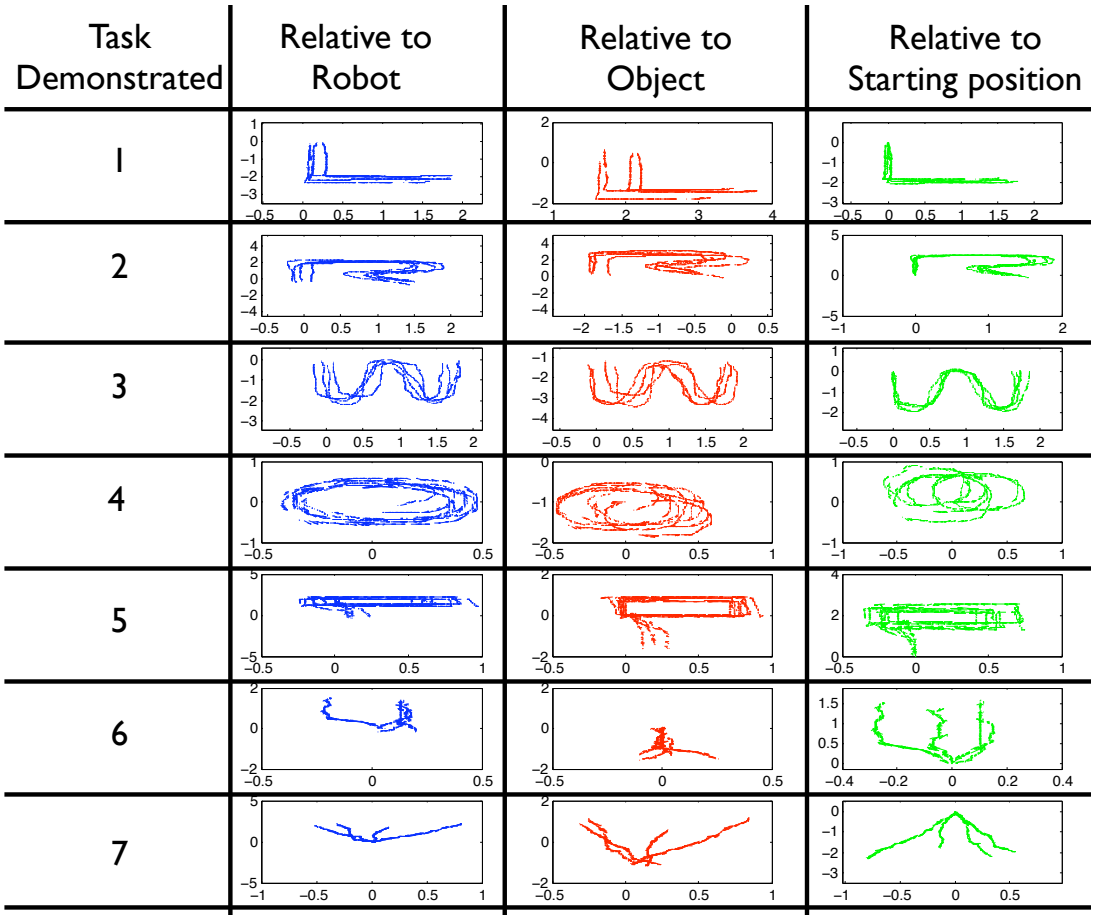

Fig. 7. The 7 motor tasks demonstrated in experiment 2 (the global contexts triggering each of them - object position, speech, gesture - are not visualized on this figure). In the second column (in blue), the demonstrated movements are represented in framing 1 (relative to the robot). In the third column (in red), the demonstrated movements are represented in framing 2 (relative to the object). And finally in the fourth column (in green), they are represented in framing 3 (relative to the starting position). The demonstrations of a task will look like several instances of a consistent policy in the correct framing but might look incoherent in the other framings. Task 1 and 2 ("L" and "R") is to be executed as a response to the object being to the left (task 1) and right (task 2). Task 3 and 4 is to be executed as a response to specific speech waves ("dubleve" and "circle" respectively). Task 5 and 6 is to be executed as a response to particular hand signs (and "S" and a "P" respectively) and task 7 is to be executed in case of a starting position far away from the robot (roughly "when the arm is extended; move close to body").

and increase with the number of irrelevant dimensions. The 6 tasks that were found all have the correct framing attached to them so when the correct task is found during reproduction the correct framing is also found. During each of the reproductions of the 6 tasks found by the grouping algorithm, the ILO-GMR algorithm was supplied by the grouping algorithm with only relevant data in only the inferred correct framing and, as can be seen in figure 10, generally performs well. Sometimes the imitator acts "twitchy" at the top of task 4 during the second time around the circle if it gets too high (this is hard to see in the figure but is apparent when watching the simulated hand move during a reproduction). The push task stops slightly to the left of the object and drifts a bit when this point is reached even if the speed is greatly reduced. The path to the object in task 6 is also not completely straight (its not straight in the demonstrations but an optimal algorithm should average the directions and smooth out these differences). The reproductions of the three tasks where framing f_s (hand position relative to the starting position) is the relevant one looks very similar since the relevant part of the starting conditions are always the same (the relevant state is position relative to the starting position so even if starting position and object position differ each time, everything that affects policy stays the same).

V. Discussion and Conclusion

Summary of results We have demonstrated that it is possible for a robotic imitator to autonomously group unlabelled demonstrations into separate tasks and to find the corresponding triggering contexts as well as the correct framings for these tasks even if the number of tasks is not provided. We have also shown that linguistic productions such as speech words or communicative gestures can be included within a generalized context, and that the imitator can determine for which tasks the linguistic part of the context is relevant.

What looks like communicative behavior to an outside observer is treated exactly the same as any other part of the context by the imitator. The fact that a single algorithm can

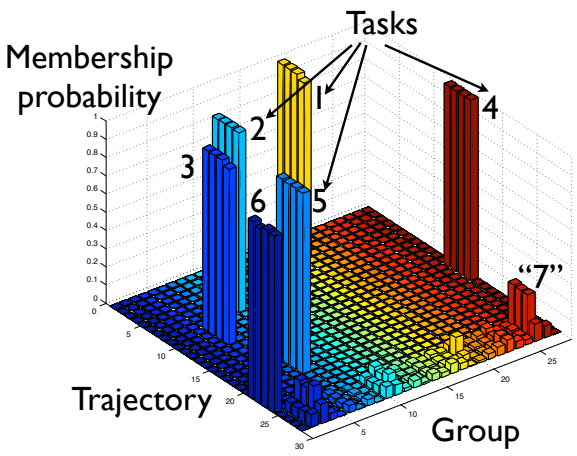

Fig. 8. The groupings found in experiment 2. The height of a pillar shows the membership value of the demonstration which number is indicated on the axis labelled "Trajectory". The 6 groups of 4 high pillars show tasks 1 to 6. There are several values on the axis labelled "Groups", representing the identities of inferred groups, that have no high values and those correspond to empty groups. For values 25 to 28 on the left demonstration axis we can see that the demonstrations of task 7 are not grouped together. The demonstrations of task number 7 have not been correctly grouped and when utilizing a cutoff value of 50% there are 6 groups formed (all of them with the correct data associated to them) but the demonstrations of task 7 are discarded (meaning that reproduction attempts of task 7 results in some other task being selected). In some runs, demonstrations 1, 2 and 3 or demonstrations 1 and 2 of task 7 are grouped together as a 7th task (which also represent a failure since while reproducing task 7, the algorithm does not have access to all the relevant information).

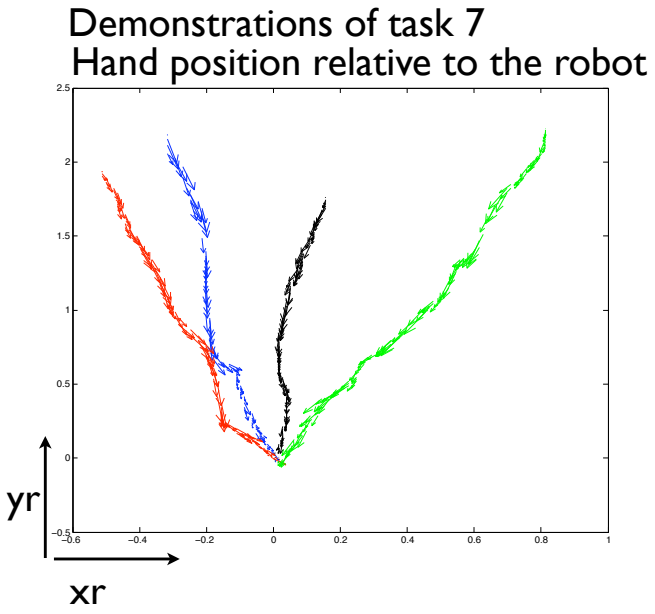

Fig. 9. Here we can see why task 7 of experiment 2 has not been correctly grouped. The similarity measure will simply not classify the red and the green demonstration as similar in any of the three framing available (they never move in the same direction, no matter how you pick points from the demonstrations). A framing that considers positions in a coordinate frame where one axis goes between the point $H_{x_r} = 0$, $H_{y_r} = 0$ and the starting point would work since the demonstrations would look the same plotted in this framing.

learn both how to respond to a traditional physical context and how to respond to a communicative act (without being told which is which and not even knowing if there is just communicative tasks, just non communicative tasks or a mix) illustrates the point that a single imitation strategy can be used for language and other sensorimotor learning.

Furthermore, experiments showed that the potential channel(s) of communication do not have to be known initially by the learner, or be confined to a single modality, but can be estimated. In the second experiment, it was also shown how some actions learnt by the learner could be described by an external observer as symbolic communicative acts (e.g. drawing a "R" when the object is to the Right, which is behaviorally like naming). Yet, for the learner, all skills are structurally similar.

Thus, the system and experiments we presented have considered *simultaneously* many kinds of ambiguities that were previously treated separately in both computational approaches to motor learning by demonstration and to language acquisition. To our knowledge, some of the ambiguities were even considered here for the first time, such as uncertainty about the communication modalities. This proof-of-concept also allowed us to show that the so-called Gavagai problem in language acquisition could be generalized, covering a family of Gavagai problems which are common to general learning of sensorimotor skills by imitation.

Limits. Yet, while the teaching data was provided by a human with a noisy movement and speech capture system, the dimensionalities of the spaces we considered here were moderate (thanks to the use of manifold projection for speech and gesture representations). This allowed us to provide visual illustrations of the result of inferences, and provide a proof-of-concept, but it remains to be evaluated to what extent such a system can be extended to more complex spaces with real world robots. Also, we did not explore the situation where a keyword for one task is spoken at the same time as a the objects position is in a region that should trigger another task. The imitator has no way of knowing how to resolve such a conflict and would need to see the teacher respond in such a situation in order to know what to do. The algorithm would pick one task based on which context is most "task typical" and execute that task as usual (the algorithm contains the relative match for the different tasks, so the information that the learner is not certain of what to do is available, but it is not currently used). Finally, there is a form of ambiguity we did not consider in this article: how to segment observations of the teacher's behavior into separate demonstrations, and how to segment demonstration into a "context" and a "response". This form of ambiguity was addressed in [44], [45], and a combination of the associated approach and the approach presented in this article would be of high interest.

Extensions and further work. In applications such as personal and assistive robotics, robots need to adapt to the preferences and particularities of each user. This implies in particular that the possibility for a robot to infer which modalities are used as communicative medium by a par-

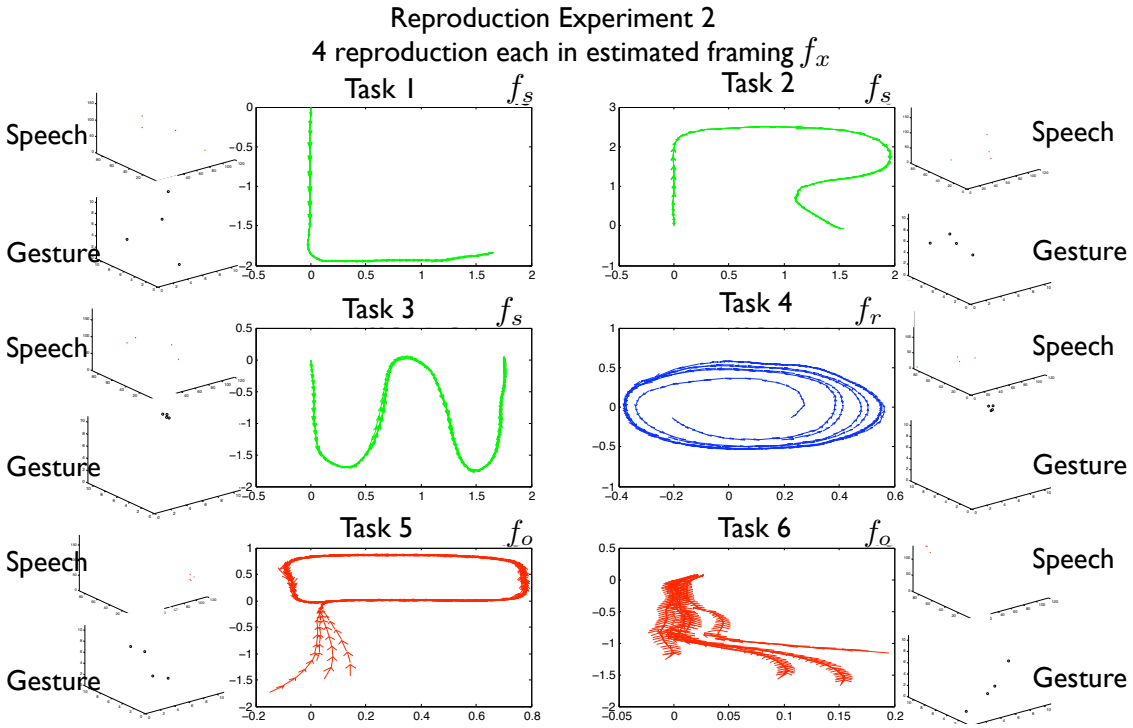

Fig. 10. **Experiment 2, test phase**: Each square shows four examples of motor responses learnt and produced by the learner in the test phase in response to global contexts corresponding to six of the tasks of experiment 2 (like during the demonstration phase, no category label is provided to the learner to indicate which task the context affords, this is inferred by the learner). Here we can see the reproductions of the 6 tasks that were correctly found by the grouping algorithm (Task 7 was not found by the grouping algorithm, and due to this, no reproduction attempts of this task were made). Each task is reproduced four times with different starting contexts (next to each square, a projection on a low-dimensional manifold of speech waves and gestures in each context is shown). The reproductions of each task is viewed in the inferred framing indicated to the top right of each subfigure. Tasks 1,2 and 3 are defined in the framing relative to the starting position of the imitator's hand, meaning that in this framing the starting position is always at $(0,0)$, resulting in more homogenous reproductions. Comparing the reproduced trajectories with the task descriptions and the demonstrations we can see that they match fairly well and comparing the inferred framings with the framings of the task descriptions we can see that all framings where inferred successfully.

ticular user, and in particular contexts, would be highly desirable. The computational architecture we presented is a first step in this direction, and further work will examine how it can allow the robot learner to infer the relevance and meaning of various other communicative social cues such as facial expressions, intonation, or gaze in addition to speech and gestures, and in the context of realistic human robot interaction.

These additional modalities, as well as the embedding of a robot learner within a social interaction loop with a non-expert human may also provide additional structure which may help the learner to identify macro-structures out of the myriad of micro-structures observable in single interactions. First, human teachers tend to spontaneously use multiple kinds of cues to disambiguate the learning situations and guide the attention of the learner towards relevant parts of the context [54], [55]: for example, they may use motherese speech or motionese to highlight novel words or important moments of a demonstrated action [56], [57], [58]. Such social attentional mechanisms could be highly helpful for a robot learner. The ability to exploit them could be pre-programmed in a robot, but an open question remains: it seems that there are important variations across learners and teachers in using these cues, so how can a learner discover them and understand their functionality before actually using them [59]? The learning architecture presented in this article explores a first step in this direction, in the sense that it shows how a learner can discover that in certain contexts, special parts of the behavior of its human peer (which an external observer could call a "social cue") become relevant and determine what other details of the context he should attend to achieve a policy.

A second crucial aspect related to the embedding of imitation learning in natural social interaction loops is the possibility to leverage mechanisms of active learning and active teaching. Indeed, human teachers continuously adapt their teaching signals to learners, and learners can trigger learning situations that provide high information gain. This can be transposed to robots, and the scalability of the unified imitation learning mechanism presented in this article would certainly be made stronger if coupled with such active mechanisms. In particular, an important extension of the work presented here would be to study how to integrate interactive learning algorithms such as presented in [23], where robots learn by demonstration and through asking questions that allow to fasten their learning process, but which were so far assuming a separate pre-programmed linguistic system to ask these questions. In order to comply with a model of early social learning which does not assume prior linguistic knowledge, such as in the context of the work we presented,

a potential route to explore is to use generic intrinsic motivation systems [60], also called curiosity-driven learning, which were already shown to self-organize the developmental discovery of early vocal interaction [61] and used to actively guide a robot learning motor skills through imitation [62]. Within such an approach, the robot learner could choose actions, as well as goals [63], that elicit a feedback with maximal information gain from the teacher, thus importantly reducing the space of possible interpretations. Such a system, combining the unified imitation learning approach presented in this article with active learning, would then constitute a useful basis to realize what has been called "teleological language and action learning" [64], where the meaning of novel actions and linguistic constructions is progressively acquired through recurrent interaction patterns, and along a process that goes from holistic partial interpretation to local compositional understanding.

Furthermore, linguistic signals considered in this article have remained at the level of lexicon: no syntax and grammar was present. Building an architecture which maintains such an homogeneity for learning in a fluid manner both non-linguistic and linguistic tasks, and which is able to detect and acquire grammatical structure, is a challenge to be addressed. A first step in this direction is presented in [2], where a variation of the architecture presented in this article allows the learner both to imitate internal cognitive operation (like attention) and to acquire the compositional meaning of two word sentences. Yet, specific mechanisms for syntax processing were included. We believe that such generic syntax processing may still be integrated uniformly in a generic architecture for sensorimotor learning by imitation, since several works have identified complex "action grammars" where syntactic operations also operate for the understanding and generation of actions [65], [66], [67].

Evolutionary hypothesis for the evolution of language. The model we presented shows that a general mechanism for learning context dependent sensorimotor skills by imitation can allow a learner to acquire simple but non-trivial linguistic skills without the addition of another mechanism. While imitation learning is only one mode of language learning among others, the strong structural similarities between action and language learning in such an imitation context suggest the hypothesis that the capability to acquire language in such a manner may be an exaptation of previously evolved capacities for general imitation learning. In such a vision, language is not only grounded in action, but language acquisition spontaneously forms out of general action learning. This may decrease the steepness of the evolutionary step from non-language to language, but emphasizes the importance of a crucial question for the origins of language: How did the capability to acquire multiple context-dependent skills through the imitation of peers, with multiple kinds of ambiguities, evolve?

REFERENCES

[1] W. V. Quine, *Word and Object*. The MIT Press, first edition ed., 1960.

[2] T. Cederborg and P.-Y. Oudeyer, "Imitating operations on internal cognitive structures for language aquisition," in *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pp. 650–657, IEEE, 2011.

[3] T. Cederborg and P.-Y. Oudeyer, "Learning words by imitating," in *Theoretical and Computational Models of Word Learning: Trends in Psychology and Artificial Intelligence* (L. Gogate and G. Hollich, eds.), pp. 296–326, IGI Global, 2013.

[4] K. Dautenhahn and C. L. Nehaniv, *Imitation in animals and artifacts*. MIT Press, 2002.

[5] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," in *Handbook of Robotics* (B. Siciliano and O. Khatib, eds.), pp. 1371–1394, Secaucus, NJ, USA: Springer, 2008.

[6] Y. Demiris and A. Meltzoff, "The robot in the crib: A developmental analysis of imitation skills in infants and robots," *Infant and Child Development*, vol. 17, no. 1, pp. 43–53, 2008.

[7] B. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.

[8] M. Lopes, F. Melo, L. Montesano, and J. Santos-Victor, "Abstraction levels for robotic imitation: Overview and computational approaches," in *From Motor Learning to Interaction Learning in Robots* (O. Sigaud and J. Peters, eds.), pp. 313–355, Springer, 2010.

[9] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 37, no. 2, pp. 286–298, 2007.

[10] S. Calinon, F. D'halluin, D. Caldwell, and A. Billard, "Handling of multiple constraints and motion alternatives in a robot programming by demonstration framework," in *Proc. IEEE-RAS Intl Conf. on Humanoid Robots (Humanoids)*, pp. 582–588, December 2009.

[11] P. Abbeel, A. Coates, and A. Y. Ng, "Autonomous helicopter aerobatics through apprenticeship learning," *Int. J. Rob. Res.*, vol. 29, pp. 1608–1639, Nov. 2010.

[12] F. Guenter, M. Hersch, S. Calinon, and A. Billard, "Reinforcement learning for imitating constrained reaching movements," *RSJ Advanced Robotics, Special Issue on Imitative Robots*, vol. 21, no. 13, pp. 1521–1544, 2007.

[13] S. Vijayakumar and S. Schaal, "Locally weighted projection regression: An o(n) algorithm for incremental real time learning in high dimensional space," in *in Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000*, pp. 1079–1086, 2000.

[14] S. Calinon, F. D'halluin, E. L. Sauser, D. G. Caldwell, and A. Billard, "Learning and reproduction of gestures by imitation," *IEEE Robot. Automat. Mag.*, vol. 17, no. 2, pp. 44–54, 2010.

[15] R. P. Rao, A. P. Shon, and A. N. Meltzoff, "A bayesian model of imitation in infants and robots," *Imitation and social learning in robots, humans, and animals*, pp. 217–247, 2004.

[16] M. Ito, K. Noda, Y. Hoshino, and J. Tani, "Dynamic and interactive generation of object handling behaviors by a small humanoid robot using a dynamic neural network model," *Neural Networks*, vol. 19, no. 3, pp. 323–337, 2006.

[17] P. Abbeel and A. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the 21st International Conference on Machine Learning*, pp. 1–8, 2004.

[18] M. Lopes, F. Melo, and J. Santos-Victor, "Abstraction levels for robotic imitation: Overview and computational approaches," in *From Motor Learning to Interaction Learning in Robots*, SpringerLink, 2010.

[19] S. Calinon, F. D'halluin, E. L. Sauser, D. G. Caldwell, and A. G. Billard, "Learning and reproduction of gestures by imitation," *Robotics & Automation Magazine, IEEE*, vol. 17, no. 2, pp. 44–54, 2010.

[20] T. Cederborg, M. Li, A. Baranes, and P.-Y. Oudeyer, "Incremental local online gaussian mixture regression for imitation learning of multiple tasks," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pp. 267–274, IEEE, 2010.

[21] O. Mangin and P.-Y. Oudeyer, "Learning the combinatorial structure of demonstrated behaviors with inverse feedback control," in *Human Behavior Understanding*, pp. 134–147, Springer, 2012.

[22] D. Kulić, C. Ott, D. Lee, J. Ishikawa, and Y. Nakamura, "Incremental learning of full body motion primitives and their sequencing through human motion observation," *The International Journal of Robotics Research*, vol. 31, no. 3, pp. 330–345, 2012.

[23] M. Cakmak and A. L. Thomaz, "Designing robot learners that ask good questions," in *Proceedings of the seventh annual ACM/IEEE*

international conference on Human-Robot Interaction, pp. 17–24, ACM, 2012.

[24] P. F. Dominey, A. Mallet, and E. Yoshida, "Progress in programming the hrp-2 humanoid using spoken language," in *Robotics and Automation, 2007 IEEE International Conference on*, pp. 2169–2174, IEEE, 2007.

[25] C. A. Tikhanoff V. and M. G, "Language understanding in humanoid robots: icub simulation experiments," in *IEEE Transactions on Autonomous Mental Development*, pp. 17–29, 2011.

[26] G. Massera, E. Tuci, T. Ferrauto, and S. Nolfi, "The facilitatory role of linguistic instructions on developing manipulation skills," *Comp. Intell. Mag.*, vol. 5, pp. 33–42, Aug. 2010.

[27] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adaptive Behavior*, vol. 13, no. 1, p. 33, 2005.

[28] L. Steels and M. Loetzsch, "Perspective alignment in spatial language," in *Spatial Language and Dialogue* (K. Coventry, T. Tenbrink, and J. Bateman, eds.), Oxford University Press, 2008.

[29] L. Steels and M. Spranger, "Can body language shape body image?," in *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems* (S. Bullock, J. Noble, R. Watson, and M. A. Bedau, eds.), pp. 577–584, MIT Press, Cambridge, MA, 2008.

[30] L. Steels and M. Spranger, "The robot in the mirror," *Connection Science*, vol. 20, no. 4, pp. 337–358, 2008.

[31] L. Steels, "Is sociality a crucial prerequisite for the emergence of language?," in *The Prehistory of Language. Oxford University Press* (R. Botha, ed.), p. 18–51, Oxford, 2008.

[32] L. Steels, "Experiments on the emergence of human communication," *Trends in Cognitive Sciences*, 2006.

[33] L. Steels, "Modeling the formation of language in embodied agents: Methods and open challenges," in *Evolution of Communication and Language in Embodied Agents* (S. Nolfi and M. Mirolli, eds.), p. 223–233, Berlin: Springer, 2010.

[34] C. Yu and D. H. Ballard, "A unified model of early word learning: Integrating statistical and social cues," *Neurocomput.*, vol. 70, pp. 2149–2165, Aug. 2007.

[35] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, and A. Zeschel, "Integration of action and language knowledge: A roadmap for developmental robotics," 2010.

[36] P.-Y. Oudeyer, *Self-Organization in the Evolution of Speech*, vol. 6 of *Studies in the Evolution of Language*. Oxford University Press, Jan. 2006.

[37] L. ten Bosch, H. V. Hamme, L. Boves, and R. K. Moore, "A computational model of language acquisition: the emergence of words," 2009.

[38] O. J. RŁsŁnen, U. K. Laine, and T. Altosaar, "Self-learning vector quantization for pattern discovery from speech.," in *INTERSPEECH*, pp. 852–855, ISCA, 2009.

[39] A. S. Park, J. R. Glass, and S. Member, "Towards unsupervised pattern discovery in speech," in *Peter Hagedorn, Wolfgang Konrad and J. Wallaschek, The Journal of Sound and Vibration*, pp. 53–58, 2005.

[40] P. Ruvolo, I. Fasel, and J. R. Movellan, "A learning approach to hierarchical feature selection and aggregation for audio classification," *Pattern Recognition Letters*, Jan. 2010.

[41] L. S, Y. E, M. A, N. F, N. L, M. G, W. F, and D. PF, "Human-robot cooperation based on learning and spoken language interaction from motor learning to interaction learning," *Robots, Studies in Computational Intelligence*, vol. 264, 2010.

[42] F. Xu, "Word learning as bayesian inference: evidence from preschoolers," in *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, pp. 288–297, Psychological, 2005.

[43] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Transactions on Applied Perception (TAP)*, vol. 1, no. 1, pp. 57–80, 2004.

[44] Y. F. O. Mohammad and T. Nishida, "Learning interaction protocols using augmented baysian networks applied to guided navigation," in *IROS*, pp. 4119–4126, 2010.

[45] Y. F. O. Mohammad, T. Nishida, and S. Okada, "Unsupervised simultaneous learning of gestures, actions and their associations for human-robot interaction," in *IROS*, pp. 2537–2544, 2009.

[46] O. Mangin and P.-Y. Oudeyer, "Learning semantic components from sub-symbolic multimodal perception," in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-Epirob)*, IEEE, 2013.

[47] M. Lopes, T. Cederborg, and P.-Y. Oudeyer, "Simultaneous acquisition of task and feedback models," in *Development and Learning (ICDL), 2011 IEEE International Conference on*, vol. 2, pp. 1–7, IEEE, 2011.

[48] J. Grizou, M. Lopes, and P.-Y. Oudeyer, "Robot learning simultaneously a task and how to interpret human instructions," in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-Epirob)*, IEEE, 2013.

[49] J. Driesen, L. ten Bosch, and H. Van Hamme, "Adaptive Non-negative Matrix Factorization in a Computational Model of Language Acquisition," in *Interspeech*, pp. 1–4, 2009.

[50] O. Mangin, P.-Y. Oudeyer, D. Filliat, *et al.*, "A bag-of-features framework for incremental learning of speech invariants in unsegmented audio streams," in *Tenth International Conference on Epigenetic Robotics*, 2010.

[51] A. Baranes and P. Y. Oudeyer, "R-IAC: Robust intrinsically motivated exploration and active learning," *Autonomous Mental Development, IEEE Transactions on*, vol. 1, pp. 155–169, Oct. 2009.

[52] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *HRI '07: Proceedings of the ACM/IEEE international conference on Human-robot interaction*, (New York, NY, USA), pp. 255–262, ACM, 2007.

[53] K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura, "Learning, generation and recognition of motions by reference-point-dependent probabilistic models," *Advanced Robotics*, vol. 25, no. 6-7, pp. 825–848, 2011.

[54] L. B. Smith, E. Colunga, and H. Yoshida, "Knowledge as process: contextually cued attention and early word learning," *Cognitive science*, vol. 34, no. 7, pp. 1287–1314, 2010.

[55] J. Saunders, C. L. Nehaniv, and C. Lyon, "The acquisition of word semantics by a humanoid robot via interaction with a human tutor," *New Frontiers in Human-Robot Interaction*, vol. 2, 2011.

[56] K. Fischer, K. Foth, K. J. Rohlfing, and B. Wrede, "Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot," *Interaction Studies*, vol. 12, no. 1, pp. 134–161, 2011.

[57] L. J. Gogate, L. E. Bahrick, and J. D. Watson, "A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures," *Child development*, vol. 71, no. 4, pp. 878–894, 2000.

[58] K. K. McGregor, K. J. Rohlfing, A. Bean, and E. Marschner, "Gesture as a support for word learning: The case of under," *Journal of child language*, vol. 36, no. 4, p. 807, 2009.

[59] K. Rohlfing and B. Wrede, "What novel scientific and technological questions does developmental robotics bring to hri?," *IEEE CIS AMD Newsletter*, vol. 9, no. 1, 2012.

[60] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 2, pp. 265–286, 2007.

[61] P.-Y. Oudeyer and F. Kaplan, "Discovering communication," *Connection Science*, vol. 18, no. 2, pp. 189–206, 2006.

[62] S. M. Nguyen and P.-Y. Oudeyer, "Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner," *Paladyn Journal of Behavioural Robotics*, vol. 3, no. 3, pp. 136–146, 2012.

[63] A. Baranes and P.-Y. Oudeyer, "Active Learning of Inverse Models with Intrinsically Motivated Goal Exploration in Robots," *Robotics and Autonomous Systems*, vol. 61, pp. 69–73, Jan. 2013.

[64] B. Wrede, K. Rohlfing, J. Steil, S. Wrede, P.-Y. Oudeyer, and J. Tani, "Towards robots with teleological action and language understanding," in *Humanoids 2012 Workshop on Developmental Robotics: Can developmental robotics yield human-like cognitive abilities?*, 2012.

[65] A. Whiten, E. Flynn, K. Brown, and T. Lee, "Imitation of hierarchical action structure by young children," *Developmental science*, vol. 9, no. 6, pp. 574–582, 2006.

[66] K. Allen, S. Ibara, A. Seymour, N. Cordova, and M. Botvinick, "Abstract structural representations of goal-directed behavior," *Psychological science*, vol. 21, no. 10, pp. 1518–1524, 2010.

[67] K. Pastra and Y. Aloimonos, "The minimalist grammar of action," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1585, pp. 103–117, 2012.

[68] S. Calinon, *Robot Programming by Demonstration: A Probabilistic Approach*. EPFL/CRC Press, 2009.

[69] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal*

Statistical Society. Series B (Methodological), vol. 39, no. 1, pp. 1–38, 1977.

VI. Appendix: Details of Learning Algorithms

A. Trajectory distance $\Delta_{t;k;i;j}$

To determine which trajectories are instances of the same movement it is necessary to define some measure of distance between two trajectories. Two trajectories that are instances of the same movement will only look similar if they are viewed in the coordinate system of the task. For this reason the distance between two trajectories is defined relative to a coordinate system. Thus $\Delta_{A;B;3}$ is the distance between trajectory A and trajectory B seen in coordinate system 3. Viewed in another coordinate system, both would still be the same type of movement, for example a circle, but they would not be centered around the same point, and thus the two trajectories would not look similar.

For each of the N points in trajectory number t the closest points in trajectory k is selected (with distance measured using the currently evaluated coordinate system). For each point p of trajectory number t, the closest point of trajectory number k is found. δ_p is defined as the angular difference in output of the two points. Then we have $D_{t;k;i} = \sum_{p=1}^{N} \delta_p^2 / N$. Finally we have $\Delta_{t;k;i} = min(D_{t;k;i}, D_{k;t;i})$.

There are many possible ways of measuring similarity between two trajectories, given the coordinate systems to view them in and the paper makes no claim on the optimality of the specific similarity measure introduced. Like many other parts of the algorithm the important part is not how the specific part is implemented but instead how it is combined with the rest of the algorithm, with the details included only for completeness.

B. Grouping algorithm

The current estimate of the probability that trajectory number t is an instance of movement number m is denoted $m_{m;t}$. The suitable value of $m_{m;t}$ is completely determined by what movements the other trajectories are estimated to be instances of. The only thing that matters is that trajectories that are instances of the same movement are grouped together. Since the number of movements is unknown there are as many movements as trajectories (so that M is a NxN matrix for N demonstrations).

Given the similarity between trajectories there are many possible ways to divide them into subgroups and the iterative algorithm proposed is not claimed to be optimal (the reader that is not interested in exactly how similarities between trajectories is used to form groups whose members have high similarity can skip this section). The basic principle of the grouping algorithm is that if two trajectories A and C are more similar to each other than other trajectories likely to be instances of movement x, then $m_{x;A}$ and $m_{x;C}$ will increase. If A and C are less similar than average, then $m_{x;A}$ and $m_{x;C}$ will decrease, and the magnitude of the change depends oh how much the similarity deviates from the other likely members.

The algorithm is described using pseudocode in 1. In order to save space, several variables (either used in the pseudocode or used to define other variables that are used in the in the pseudocode) are defined and explained below rather than in the pseudocode, such as: maximum trajectory similarity $\gamma_{t;k}$, joint memberships: $\omega_{t;k}$, weighted mean similarity ϖ_t and push strength $\xi_{t;k}$.

Maximum trajectory similarity $\gamma_{t;k}$. $\gamma_{t;k;i}$ is the inverse of the distance $\Delta_{t;k;i}$ and $\gamma_{t;k}$ is the maximum similarity between trajectories t and k, $\gamma_{t;k} = max_i(\gamma_{t;k;i})$ (for example, if trajectories A and C have the highest similarity when in coordinate system 1, then $\gamma_{A;C} = \gamma_{A;C;1}$.

Joint memberships $\omega_{t;k}$ is a measure of how probable it is that trajectories t and k are instances of the same movement according to the current state of the membership matrix M. It is calculated as: $\omega_{t;k} = (max_m(m_{m;t} * m_{m;k}))/(\sum_{\tau=1}^{N} max_{m;\tau}(m_{m;t} * m_{m;\tau}))$.

Weighted mean similarity ϖ_t is a measure of the weighted average similarity to trajectory t of trajectories that are likely to be instances of the same movement. $\varpi_t = \sum_{k=1}^{N} \omega_{t;k} * \gamma_{t;k}$.

Push strength $\xi_{t;k}$ is the strength with which trajectory t will affect the memberships of trajectory k in the movement groups that they are both probable members of. If it is positive the presence of trajectory k in a movement group will increase the membership of trajectory t and decrease it if it is negative. It is calculated as: $\xi_{t;k} = e^{((\gamma_{t;k}/\varpi_t)-1)}$, and we can for example see that $\xi_{t;k} = 1$ if the similarity between t and k is exactly the same as the average weighted similarity between t and the other trajectories that has high joint memberships with t. If the similarity $\gamma_{t;k}$ is bigger than the weighted average ϖ_t, the we will get a push strength $\xi_{t;k} > 1$ (and if the similarity $\gamma_{t;k}$ is smaller than the weighted average ϖ_t, we will get $\xi_{t;k} < 1$).

C. Incremental Local Online Gaussian Mixture Regression (ILO-GMR)

Both experiments uses the ILO-GMR regression approach which here takes as input demonstrations of the task that is to be performed, as well as information about what task space/framing to use, and outputs actions. It is a modification of the GMR method, which has already been well explored in the context of imitation learning (see for example an experiment [9], a book with focus on GMR [68] or an experiment combining GMR with HMM and learning two tasks from unlabelled demonstrations [14]).

1) Gaussian Mixture Regression (GMR):: The GMR approach first builds a model using a Gaussian Mixture Model encoding the covariance relations between different variables. If the correlations vary significantly between regions then each local region of state space visited during the demonstrations will need a few gaussians to encode this local dynamics. Given data and the number of gaussians, the use of an Expectation Maximization (EM) algorithm [69] finds the parameters of the model.

A Gaussian probability density function consists of a mean μ and a covariance matrix Σ. The probability density ρ of observing the output v from a gaussian with parameters μ and Σ is:

$$\rho(v) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\{-\frac{1}{2}(v-\mu)^T \Sigma^{-1}(v-\mu)\} \qquad (1)$$

Algorithm 1 Overview of the iterative grouping algorithm

Input: M_1, S, N
- M_1 *is the initial membership probabilities*
- S *is the number of steps (S=50 is used here)*
- N *is the number of demonstrations*

for $s = 1$ **to** S **do**
  $M_{mod} \leftarrow M_s$ ($m_{m;t}$ refers to M_{mod})
  $M_{old} \leftarrow M_s$ ($m_{m;t;old}$ refers to M_{old})
  for $m = 1$ **to** N **do**
    for $t = 1$ **to** N **do**
      for $k = 1$ **to** N, $k \neq t$ **do**
        $m_{m;t} \leftarrow m_{m;k;old}\xi_{k;t} + (1 - m_{m;k;old})m_{m;t}$
      end for
    end for
  end for
  Rescale
  Preferring hypotheses with few movement types:
  $\forall : 1 < m < N, 1 < t < N$**:**
  $m_{m;t} \leftarrow m_{m;t} \times (\sum_{\tau=1}^{N} m_{m;\tau})^{1/4}$
  Rescale
  $m_{m;t} \leftarrow m_{m;t} + 0.0001$
  Rescale
  $M_{s+1} \leftarrow M_{mod}$
end for
note that if the push factor $\xi_{t;k}$ is positive $m_{m;t}$ will increase and if it is negative it will decrease in the central update step. Remember that a positive $\xi_{t;k}$ indicates that the policy similarity between t and k is higher than the weighted average. The rescaling makes the memberships of a single demonstration sum to 1

To get the best guess of the desired output $\hat{v}$ (e.g. speed in cartesian space of the hand, as in the experiments presented here) given only the current state x_q (e.g. position and speed of the hand in various referentials and position of an object construing the context, as in the experiments presented here) we have:

$$\hat{v}(x_q) = E[v|x = x_q] = \mu^v + \Sigma^{vx}(\Sigma^{xx})^{-1}(x_q - \mu^x) \quad (2)$$

Where Σ^{vx} is the covariance matrix describing the covariance relations between x and v.

A single such density function can not encode non linear correlations between the different variables. To do this we need to use more than one gaussian to form a Gaussian Mixture Model defined by a parameter list $\lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_M\}$, where $\lambda_i = (\mu_i, \Sigma_i, \alpha_i)$ and α_i is the weight of gaussian i. To get the best guess $\hat{v}$ conditioned on an observed value x_q we first need to know the probability $h_i(x_q)$ that gaussian i produced x_q. This is simply the density of the gaussian i at x_q divided by the sum of the other densities at x_q, $h_i(x_q) = \frac{\rho_i(x_q)}{\sum_{j=1}^{M} \rho_j(x_q)}$ (where each density $\rho_i(v)$ is calculated just as in (1), with Σ replaced by Σ_i^{xx}, v with x_q, etc). Writing out the whole computation we have:

$$h_i(x_q) = \frac{\frac{\alpha_i}{\sqrt{|\Sigma_i^{xx}|}} \exp\{-\frac{1}{2}(x_q - \mu_i^x)^T(\Sigma_i^{xx})^{-1}(x_q - \mu_i^x)\}}{\sum_{j=1}^{M} \frac{\alpha_j}{\sqrt{|\Sigma_j^{xx}|}} \exp\{-\frac{1}{2}(x_q - \mu_j^x)^T(\Sigma_j^{xx})^{-1}(x_q - \mu_j^x)\}}. \quad (3)$$

Given the best guesses $\hat{v}_i(x_q)$ from (2), and the probabilities $h_i(x_q)$ that gaussian i generated the output, the best guess $\hat{v}(x_q)$ is given by:

$$\hat{v}(x_q) = \sum_{i=1}^{M} h_i(x_q)\hat{v}_i(x_q) \quad (4)$$

The parameter list is found using an Expectation Maximization algorithm (EM) [69] that takes as input the number of gaussians and a database.

a) ILO-GMR: In these experiments the algorithm takes selected demonstrations as inputs (assuming that they have been grouped by the grouping algorithm above and thus that they are all of the same task). The datapoints of all those demonstrations are stored in D. Then, during each iteration of the reproduction of a task the imitator looks at its current state x_q and extracts a local database $D(x_q)$ consisting of the N points closest to x_q (measuring distance in the task space). These points are now used as input to GMR as described above. N is the first parameter of ILO-GMR and is typically slightly superior to the second parameter M multiplied by the dimensionality of the sensorimotor space. The EM algorithm builds a GMM and then we get the best guess of the current desired speed $\hat{v}(x_q, D(x_q), N, M)$ as described above. So at each iteration new local data is extracted and a new local model is built and used to find the desired direction.

ILO-GMR was previously used to learn four different sensorimotor tasks simultaneously [20], where the task that should be performed was only dependent on the location of an object, and was shown to perform at least as well as state-of-the-art regression methods for learning high-dimensional robot forward models, while being much easier to tune [51].