A framework integrating statistical and social cues to teach a humanoid robot new skills

Sylvain Calinon and Aude Billard

***Abstract*— Bringing robots as collaborative partners into homes and offices presents various challenges to human-robot interaction. Robots will need to interact with untrained users in environments that are originally designed for humans. Compared to their industrial homologous form, humanoid robots can not be preprogrammed with an initial set of behaviours. They should adapt their skills to a huge range of possible tasks without needing to change the environments and tools to fit their needs. The rise of these humanoids implies an inherent social dimension to this technology, where the end-users should be able to teach new skills to these robots in an intuitive manner, relying only on their experience in teaching new skills to other human partners. In previous work, we developed a generic Robot Programming by Demonstration (RPD) framework to extract the task constraints from cross-situational observations. In this paper, we present our ongoing research towards integrating information from various social cues such as joint attention or vocal intonation to this probabilistic framework.**

I. INTRODUCTION

For an efficient collaboration with human users, indoor robots such as humanoids should be provided with adaptive controllers that can behave robustly in changing situations. These robots should be provided with natural interfaces to interact easily and naturally with end-users [1], and it should be possible to reprogram them in an intuitive manner [2]. Indeed, as these robots are supposed to use a very wide range of infrastructures and tools designed originally for humans, it is not possible to pre-encode all the gestures that will be required to perform skills such as manipulation tasks. It is therefore crucial to facilitate the skill transfer process by providing end-users with natural teaching methods to reprogram these robots in an intuitive manner.

Robot Programming by Demonstration (RPD) covers such methods by which a robot learns new skills through human guidance. This paper presents our ongoing research towards bringing user-friendly human-robot teaching systems that would speed up the skill transfer process. To do so, we suggest to use a generic probabilistic framework gathering information from cross-situational observations of a skill with information extracted from different social cues observed during the interaction. Fig. 1 presents the architecture of the proposed framework.

This work was supported by the European Commission as part of the Robot@CWE project (`http://www.robot-at-cwe.eu`) under contract FP6-2005-IST-5, and as part of the FEELIX GROWING project (`http://www.feelix-growing.org`) under contract FP6 IST-045169.

S. Calinon and A. Billard are with the Learning Algorithms and Systems Laboratory (LASA), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland `{sylvain.calinon,aude.billard}@epfl.ch`

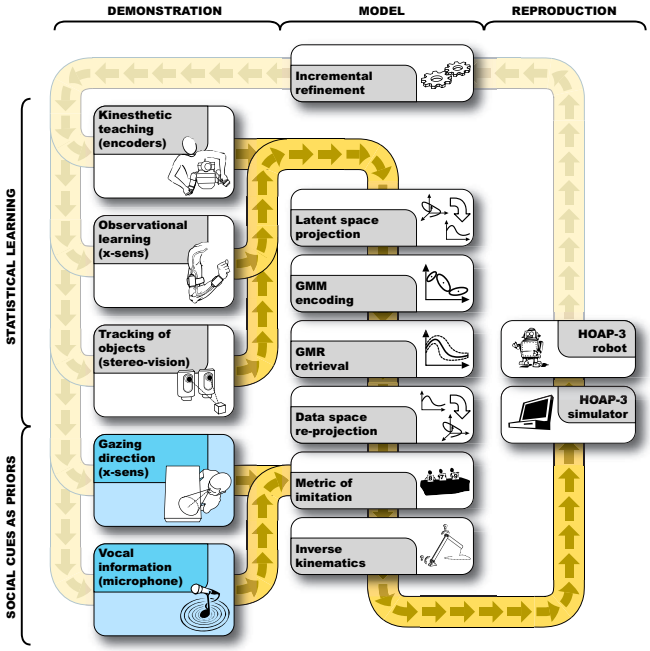

Fig. 1. Information flow across the complete system, where the constraints of a task are extracted through multiple demonstrations performed in slightly different situations and by using various social cues to scaffold the teaching interaction.

A. Robot programming by demonstration

Generic approaches to transfer new skills to a robot are those that allow the robot to extract automatically what are the important features characterizing each task and to search for a controller that optimizes the reproduction of these characteristic features. A key concept at the bottom of these approaches is that of determining a *metric of imitation performance*. One must first determine the metric, i.e. determine the weights one must attach to reproducing each of the components of the skill. It is then possible to find an optimal controller for imitation by trying to minimize this metric (e.g., by evaluating several reproduction attempts or by deriving the metric to find an optimum). The metric acts as a cost function for the reproduction of the skill [3]. In other terms, a metric of imitation provides a way of expressing quantitatively the user's intentions during the demonstrations and to evaluate the robot's faithfulness at reproducing those. To learn the metric (i.e. infer the task constraints), one common approach consists of creating a model of the skill based on several demonstrations performed in slightly different conditions (cross-situational statistical

learning). This generalization process consists of exploiting the variability inherent to the various demonstrations to extract which are the essential components of the task. These essential components should be those that remain invariant across the various demonstrations.

A large body of work explored the use of a symbolic representation to both the learning and the encoding of skills and tasks, see e.g. [4], [5]. The main advantage of a symbolic approach is that high-level skills (consisting of sequences or hierarchies of symbolic cues) can be learned efficiently through an interactive process. However, because of the symbolic nature of their encoding, these methods rely on a large amount of prior knowledge to predefine the important cues and to segment those efficiently.

Another body of work focusses on representing the task constraints at a trajectory level to avoid putting too much prior knowledge in the controllers required to reproduce a skill, see e.g. [6], [7].[1] We follow this approach in our work by using *Gaussian Mixture Model* (GMM) and *Gaussian Mixture Regression* (GMR) to respectively encode a set of trajectories and retrieve a smooth generalized version of these trajectories and associated variabilities.

The remainder of this paper is organized as follows. Section II presents the statistical learning framework used to encode the skill (II-A), showing how to generalize a learned task to various situations by considering several constraints (II-B), and showing how different modalities can be used to demonstrate a skill (II-C). Section III then illustrates how the statistical learning approach can be enhanced by social cues such as the orientation of the head while demonstrating a manipulation skill involving objects (III-A) or the variations of intonation in the vocal trace to bring the attention of the robot to particular events while demonstrating the skill (III-B). Section IV discusses the results and presents further work.

II. EXTRACTING TASK CONSTRAINTS THROUGH STATISTICAL LEARNING

A. Encoding and generalization

Through the use of *Gaussian Mixture Model* (GMM), we showed in previous work that a robot could extract autonomously the essential characteristics of a set of trajectories captured through the demonstrations [10], and that *Gaussian Mixture Regression* (GMR) could be used to retrieve a generalized version of the trajectories either in joint space [11], or in task space [12]. Table I presents the procedure for the encoding and generalization of the skill. The optimal number of components is estimated here through *Bayesian Information Criterion* (BIC) [13].

B. Reproduction by considering multiple constraints

To find a controller for the robot that takes into account constraints both in joint space and in task space, as well as the kinematic redundancy of the humanoid arm, we proposed two inverse kinematics (IK) approaches: (1) a method based on Jacobian computation using Lagrange optimization which allows to handle constraints on multiple objects in task space and in joint space simultaneously [10]; and (2) a geometric inverse kinematics approach for a 4 DOFs humanoid arm, by representing the motion of the arm by the 3D Cartesian path of the hand and by an additional parameter representing the elevation of the elbow with respect to a vertical plane [12]. Here, the geometric inverse kinematics method is used as it is much simpler for the 4 DOFs arm considered.

We illustrate the generalization and reproduction methods

[1]For an exhaustive review and comparisons of the different methods proposed in RPD, the interested reader can refer to [2], [8].

TABLE I

PROBABILISTIC ENCODING OF THE TASK CONSTRAINTS AND GENERALIZATION THROUGH GAUSSIAN MIXTURE REGRESSION (GMR).

- The dataset $x = \{x_j\}_{j=1}^{N}$ is defined by N observations $x_j \in \mathbb{R}^D$ of sensory data changing through time, where each demonstration is rescaled to a fixed duration T. Each datapoint $x_j = \{t_j, x_j^S\}$ consists of a temporal value $t_j \in \mathbb{R}$ and a spatial vector $x_j^S \in \mathbb{R}^{(D-1)}$.
- The dataset x is first modelled by a *Gaussian Mixture Model* (GMM) of K components. Each datapoint x_j is then defined by its probability density function

$$p(x_j) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x_j; \mu_k, \Sigma_k),$$

  where π_k are prior probabilities and $\mathcal{N}(\mu_k, \Sigma_k)$ are Gaussian distributions defined by centers μ_k and covariance matrices Σ_k, whose temporal and spatial components can be represented separately as

$$\mu_k = \{\mu_k^T, \mu_k^S\}, \quad \Sigma_k = \begin{pmatrix} \Sigma_k^{TT} & \Sigma_k^{TS} \\ \Sigma_k^{ST} & \Sigma_k^{SS} \end{pmatrix}.$$

- For each component k, the expected distribution of x_j^S given the temporal value t_j is defined by

$$\begin{aligned} p(x_j^S | t_j, k) &= \mathcal{N}(x_j^S; \hat{x}_k^S, \hat{\Sigma}_k^{SS}), \\ \hat{x}_k^S &= \mu_k^S + \Sigma_k^{ST}(\Sigma_k^{TT})^{-1}(t_j - \mu_k^T), \\ \hat{\Sigma}_k^{SS} &= \Sigma_k^{SS} - \Sigma_k^{ST}(\Sigma_k^{TT})^{-1}\Sigma_k^{TS}. \end{aligned}$$

- By considering the complete GMM, the expected distribution is defined by

$$p(x_j^S | t_j) = \sum_{k=1}^{K} \beta_{k,j} \, \mathcal{N}(x_j^S; \hat{x}_k^S, \hat{\Sigma}_k^{SS}),$$

  where $\beta_{k,j} = p(k|t_j)$ is the probability of the component k to be responsible for t_j, i.e.,

$$\beta_{k,j} = \frac{p(k)p(t_j|k)}{\sum_{i=1}^{K} p(i)p(t_j|i)} = \frac{\pi_k \mathcal{N}(t_j; \mu_k^T, \Sigma_k^{TT})}{\sum_{i=1}^{K} \pi_i \mathcal{N}(t_j; \mu_i^T, \Sigma_i^{TT})}.$$

- By using the linear transformation property of Gaussian distributions, an estimation of the conditional expectation of x_j^S given t_j is thus defined by $p(x_j^S | t_j) \sim \mathcal{N}(\hat{x}_j^S, \hat{\Sigma}_j^{SS})$, where the parameters of the Gaussian distribution are defined by

$$\hat{x}_j^S = \sum_{k=1}^{K} \beta_{k,j} \, \hat{x}_k^S, \quad \hat{\Sigma}_j^{SS} = \sum_{k=1}^{K} \beta_{k,j}^2 \, \hat{\Sigma}_k^{SS}.$$

- By evaluating $\{\hat{x}_j^S, \hat{\Sigma}_j^{SS}\}$ at different time steps $t_j \in [0, T]$, a generalized form of the trajectories $\hat{x} = \{t_j, \hat{x}_j^S\}$ and associated covariance matrices $\hat{\Sigma} = \{\hat{\Sigma}_j^{SS}\}$ representing the constraints along the task can then be computed (see also [9]).

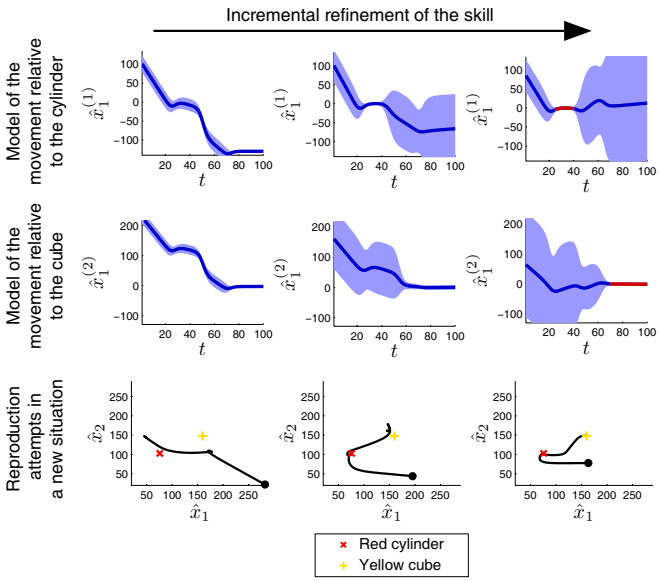

Fig. 2. Incremental refinement of the task depicted in Fig. 3, coded in a frame of reference located on the objects that are manipulated (only a subset of the variables are shown, see Fig. 3 for the frame of reference). The three columns correspond respectively to a representation of the task constraints after 1, 3 and 6 demonstrations. The first two rows show the refinement of the *Gaussian Mixture Regression* (GMR) model representing the constraints for the cylinder (first row) and for the cube (second row) along the movement. After a few demonstrations, we see that the trajectories relative to the two objects are highly constrained for particular subparts of the task, namely when reaching for the cylinder (thin envelope around time step 30) and when placing it on top of the cube (thin envelope around time step 100). The last row shows the robot's reproduction attempts (after 1, 3 and 6 demonstrations) for a new situation that has not been demonstrated. We see that after 6 demonstrations, the robot correctly reproduces the essential characteristics of the skill, namely reaching for the cylinder and dropping it on the cube (see [12] for a complete description of the results).

with an experiment involving manipulation and displacement of objects. In this experiment, the skill is represented as constraints in task space by considering the right hand path relative to two objects observed by the robot in its environment. The constraints associated with the position of the right hand with respect to an object n are thus represented by the generalized trajectory $\hat{x}^{(n)}$ and associated covariance matrices $\hat{\Sigma}^{(n)}$ (see Table I).

Fig. 2 shows how GMR encapsulates the task constraints through the generalization process. Fig. 3 shows the results of the generalization process (after six demonstrations) through snapshots of the robot reproducing the learned skill in a new situation (new initial positions of objects).

C. Incremental refinement of the skill, use of different modalities and scaffolding process

A trend of research draws the attention on the role of the teacher as being one of the most important key component for an efficient transfer of the skill, where the teaching interaction allows the user to become an active participant in the learning process (and not only a model of expert behaviour), see e.g. [12], [14]–[18]. This active teaching process allows the learner to experience and adapt the skill for his/her particular body capacities, as suggested by

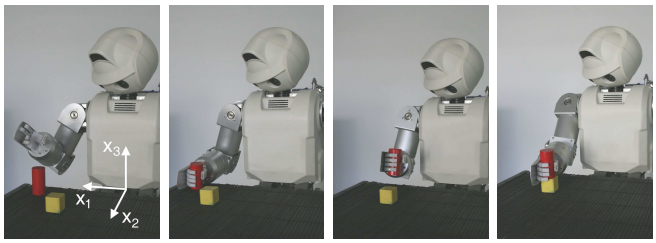

Fig. 3. Example of a manipulation task using two objects, where constraints on the hand-objects relationships along the motion are extracted probabilistically, namely grasping the red cylinder, reaching for the yellow cube (by using a bell-shaped trajectory to avoid hitting the cube), and dropping the cylinder on top of the cube. The statistical representation of the task constraints then allows the robot to reproduce the skill with different initial positions of the objects. For this experiment, a Fujitsu HOAP-3 humanoid robot with 4 DOFs for the right arm and 1 DOF for the hand is used.

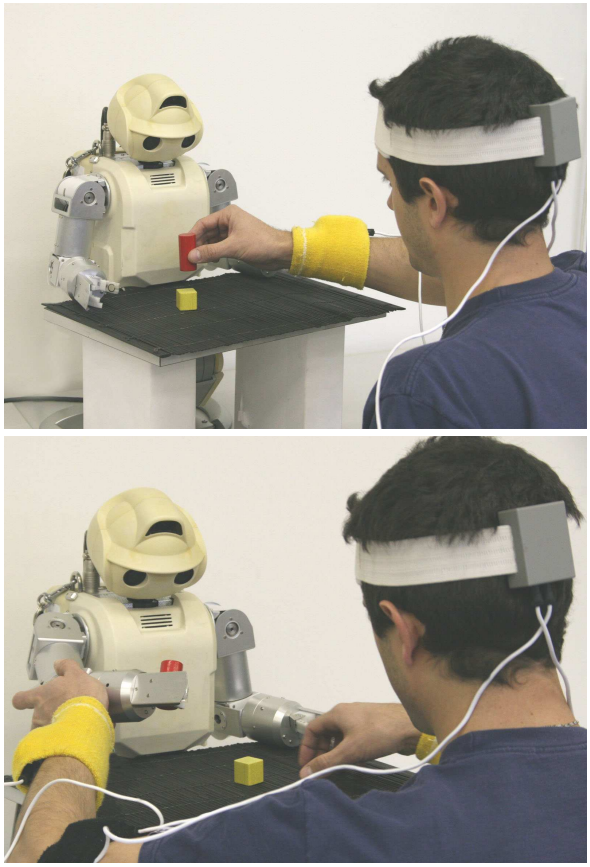

Fig. 4. Different modalities are used to convey the demonstrations and scaffolds required by the robot to learn a skill. The user first demonstrates the whole movement while wearing motion sensors (*top*) and then helps the robot refine its skill through kinesthetic teaching (*bottom*), that is, by grasping the robot's arms and moving them through the motion. 4 *X-Sens* motion sensors attached to the torso, right upper-arm, right lower-arm, and back of the head are first used to decompose the 3D absolute orientation of each segment into a set of joint angles. Through direct kinematics, the position of the hand in the 3D Cartesian space is then estimated. For kinesthetic teaching, the motor encoders of the robot are used to record information while the teacher moves the robot's arms. The user first selects the motors to control manually by slightly moving the corresponding limbs just a few milliseconds before the reproduction starts. The selected motors are set to passive mode, which allows the user to move freely the corresponding degrees of freedom while the robot executes the task, thus providing partial demonstrations while the robot executes the remaining motion. For these two methods, two webcams within the robot's head are used to track the 3D position of the objects (see [12] for details).

developmental psychology studies [19].

Following this approach, Riley *et al* [17] highlighted the importance of an active participation of the teacher not only to demonstrate a model of expert behaviour but also to refine the acquired motion by vocal feedback. Saunders *et al* [4] provided experiments where a wheeled robot is teleoperated through a screen interface to simulate a *moulding* process, that is, by letting the robot experience sensory information when exploring its environment through the teacher's support. Rohlfing *et al* highlighted the importance of having multimodal cues to reduce the complexity of human-robot skill transfer [18]. In their work, they consider multimodal information as an essential element to structure the demonstrated tasks. Through experiments, they showed that humans transfer their knowledge in a social interaction by recognizing what current knowledge the learner lacks. They then suggested taking insights from these studies to reduce the learning complexity of current RPD frameworks. Thus, sharing human adaptability with the less knowledgeable becomes a central issue when designing social robots, and they therefore hypothesize that a human teacher can also adapt naturally to a robot equipped with specific abilities.

In [12], we adopted a similar strategy and showed that the skill transfer process can benefit from the user's capacity to adapt his/her teaching strategies to the particular context. We extended the concept to the learning of continuous motion trajectories and of actions on objects, and proposed experiments where a humanoid robot learns new manipulation skills by first observing a human demonstrator (through motion sensors) and then gradually refining its skill through kinesthetic teaching (see Fig. 4). In this application, the user provides scaffolds to the robot for the reproduction of the skill by moving kinesthetically a subset of the motors. Through the supervision of the user who progressively dismantles the scaffolds after each reproduction attempt, the robot can finally reproduce the skill on its own (see also Figs 3 and 2).

We thus suggest to use different modalities to produce the demonstrations, similarly to a teaching process where a human teacher would first demonstrate the complete skill to the learner, followed by practice trials performed by the learner under the supervision of the teacher. We take the perspective that unlike observational learning, *pedagogy* is required to facilitate the transfer of the skill, which is a special type of communication used to manifest the relevant knowledge of a skill.

As discussed by Gergely and Csibra, the teacher first needs to analyze his/her knowledge content to emphasize in his/her demonstrations the aspects that are relevant for the learner [20]. In our experiments, observational learning is used similarly as a first method for the user to demonstrate natural gestures by controlling simultaneously a large number of degrees of freedom (up to 14 joint angles for the tasks considered in our experiments). Kinesthetic teaching then provides a way of supporting the robot in its reproduction of the task. Through this scaffolding process, the user provides support to the robot by manually articulating a decreasing

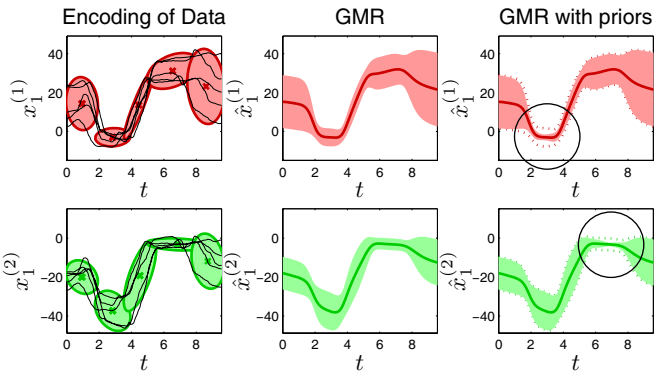

Fig. 5. Influence of the speech/gaze priors on constraints extracted through statistical learning for an interaction with a naïve user. The first and second rows correspond to the trajectories relative respectively to the cylinder and to the cube (see Fig. 3). *First column:* Data extracted from 5 demonstrations and encoding through a *Gaussian Mixture Model* (GMM) of 5 components (the optimal number of components is estimated through Bayesian Information Criterion (BIC) [13]). *Second column:* Generalization of the trajectories through *Gaussian Mixture Regression* (GMR) based solely on cross-situational statistics. In this case, we see that five demonstrations are not sufficient to extract interesting information concerning the task constraints, i.e. the envelope thickness around the generalized trajectory does not present much variations (see Fig. 2 for comparison). *Third column:* By using GMR with the social priors defined in Eqs (1), (2) and (3), we see that the envelopes become thinner in the relevant parts of the trajectories, namely when grasping the cylinder and when dropping it on the cube (highlighted by two circles in the graphs).

subset of motors. The scaffolds progressively fade away and the user finally lets the robot perform the task on its own, allowing the robot to experience the skill independently.

One advantage of this approach is that the user can provide partial demonstrations by using the robot's own kinematics and can demonstrate the task in the robot's own environment. This kinesthetic teaching process also allows the user to feel the robot's body limitations and provide appropriate examples that take these limitations into consideration.

To apply technically this teaching approach, we also demonstrated in [11] that it was possible to use a GMM/GMR framework to learn a skill incrementally and in an on-line manner without having to keep each demonstration in memory. Such an incremental learning approach allows the teacher to watch the robot's reproduction attempts after each demonstration, and thus helps him/her assess the robot's current understanding of the skill and prepare the following demonstration accordingly.

III. EXTRACTING TASK CONSTRAINTS THROUGH SOCIAL CUES

The system presented in the previous section requires to observe the skill in slightly different situations. Even if this variation appears naturally when executing the skill several times, the robot's capacity to generalize over different contexts also depends on the pedagogical quality of the demonstrations provided (e.g. gradual variability of the situations and exaggerations of the key features to reproduce).

This fact shares similarities with the human way of teaching. Indeed, a good teacher also extends the demonstrations progressively so that the learner can more easily

infer the connections between the different examples, and the range of the possible situations where the skill may apply is progressively increased. Thus, by using a statistical learning strategy alone, we implicitly suggest that one way of increasing the speed of the teaching process is to rely on the user's natural propensity for teaching by structuring the successive examples provided and guiding the learner's exploration.

Indeed, in our teaching scenarios up until now, an expert user displaces progressively the objects after each demonstration to provide variability in the exposures of the skill. In such a situation, it is nearly always possible for the robot to extract the task constraints with only a few demonstrations (from four to ten for most of the tasks that we have considered). However, it may happen that untrained users provide a set of demonstrations remaining either too similar or too different from one example to the other. In this case, a larger set of demonstrations would be required to generalize the skill. The first two columns of Fig. 5 show an experiment similar to the one performed by an expert user (presented in Fig. 2), where the untrained user provided five demonstrations that were too similar to extract correctly the task constraints through statistics.

To weaken the drawback of such situations, we follow a learning approach where the joint use of cross-situational observations and social cues ensures an efficient transfer of the task through interaction with the user. We thus propose to enhance the statistical learning strategy with information coming from various social cues, and show that these cues can be represented statistically as priors in the GMM/GMR framework. We focus here on gaze and speech information to demonstrate that interactional cues of different natures can be considered. It is important to note that we do not aim at developing state-of-the-art gaze tracking systems or speech recognition systems. We only describe here prototypes of these systems to show that multimodal social cues can be integrated in our framework through generic probabilistic approaches.

By using a computer game, Thomaz and Breazeal explored the ways in which machine learning can be designed to take advantages of natural human interaction and tutelage. They demonstrated that augmenting a reinforcement learning process with the social mechanisms of attention direction and gaze positively impacts the dynamics of the underlying learning mechanisms, highlighting the reciprocal nature of the teaching-learning partnership [15]. They showed through their experiments that the teacher's ability to guide the learner's attention to appropriate objects at appropriate times creates a significantly more robust and efficient learning interaction. Our work follows a similar approach and extends this concept to the learning of continuous gestures by a humanoid robot.

In the field of speech acquisition and word learning, Yu and Ballard explored how humans can learn words/objects couplings through statistical learning, and proposed a model for early word acquisition in a unified framework integrating statistical and social cues [21]. Their model links these two

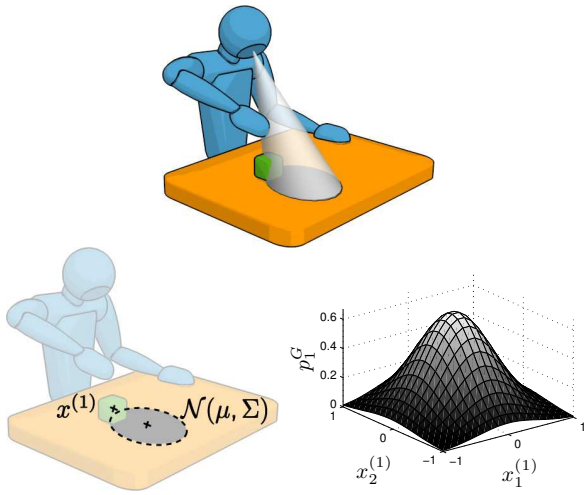

Fig. 6. Illustration of the use of gaze information to speed up the learning process through a probabilistic measure of saliency when the head is turned toward an object (see also Table II). *Top:* Estimation of gaze direction by representing the position and orientation of the head (extracted through *X-Sens* motion sensors) as a cone of vision which intersects with a surface. *Bottom-left:* Probabilistic representation of the intersection as a 2D Gaussian distribution. *Bottom-right:* Estimation of the probability to focus the attention of the robot at a particular time on a particular object i placed on the table knowing its initial position $x^{(i)}$ (the object position is tracked through the robot's built-in stereoscopic vision system).

sources of information by considering joint attention and prosody on the one hand, and statistics from cross-situational observations on the other hand. Our work follows a similar approach by extending the concept to robot learning by imitation.

A. Use of head/gaze information as priors

A large body of work explored the use of gaze direction and head orientation as a way to convey the intention of the user, see e.g. [14], [22], [23]. In [24], we showed that we can roughly extract gaze information by measuring the orientation of the head through *X-Sens* motion sensors. Even if this remains a strong assumption (as head orientation can not be considered directly as a social cue), it affects gaze following, i.e., the head naturally turns towards a goal when there is no other constraint.

Fig. 6 and Table II present a method to detect joint attention by representing gaze direction as a cone of vision whose intersection with a table can be represented as a Gaussian distribution. Instead of simply defining an attention point as the intersection of a gaze direction line with a plane, this method also evaluates the robustness of the measure through a covariance matrix. Following this method, the last row of Fig. 7 presents an example of the probability p^G extracted along the task, representing the probability of bringing the robot's attention to one of the objects detected in the scene (either the cylinder or the cube).

B. Use of vocal information as priors

Vocal deixis using speech recognition engines has been explored as a way for the user to highlight through linguistic information the steps of the demonstration that are deemed

TABLE II

PROBABILISTIC ESTIMATION OF THE GAZE DIRECTION.

- The gaze is modeled by a cone of vision with vertex point t_1, direction d_1 and half-cone angle θ. A point x on the cone satisfies

$$d_1 \left(\frac{x - t_1}{|x - t_1|} \right) = \cos(\theta),$$

or in a matrix form (I denotes the identity matrix)

$$\begin{aligned} (x - t_1)^T M (x - t_1) &= 0, \\ \text{with} \quad M &= d_1 d_1^T - (\cos(\theta))^2 \, I. \end{aligned}$$

- The table is defined by a plane with origin t_2, and directions d_{21} and d_{22}. A point x on the plane satisfies

$$x = t_2 + x_1 \, d_{21} + x_2 \, d_{22}.$$

- The intersection of the cone and the plane defines a conic

$$c_1 \, x_1^2 + 2c_2 \, x_1 x_2 + c_3 \, x_2^2 + 2c_4 \, x_1 + 2c_5 \, x_2 + c_6 = 0,$$

with $t_{12} = t_2 - t_1$, $c_1 = d_{21}^T M d_{21}$, $c_2 = d_{21}^T M d_{22}$, $c_3 = d_{22}^T M d_{22}$, $c_4 = t_{12}^T M d_{21}$, $c_5 = t_{12}^T M d_{22}$ and $c_6 = t_{12}^T M t_{12}$.

- This conic can be re-written in an homogenous matrix form

$$x^T C x = 0 \, , \quad C = \begin{pmatrix} c_1 & c_2 & c_4 \\ c_2 & c_3 & c_5 \\ c_4 & c_5 & c_6 \end{pmatrix} = \begin{pmatrix} C_R & C_t \\ C_t^T & C_\delta \end{pmatrix},$$

where $x = (x_1, x_2, 1)^T$, $C_R \in \mathbb{R}^{2\times 2}$, $C_t \in \mathbb{R}^{1\times 2}$ and $C_\delta \in \mathbb{R}$.

- The canonical form of the conic C_c is determined by transforming the conic matrix C through an Euclidean transformation H

$$\begin{aligned} C_c = \begin{pmatrix} C_{c1} & 0 & 0 \\ 0 & C_{c2} & 0 \\ 0 & 0 & C_{c3} \end{pmatrix} &= H^T C H, \\ \text{with} \quad H &= \begin{pmatrix} R & t \\ 0^T & 1 \end{pmatrix}, \end{aligned}$$

where the rotation R and translation t are found by diagonalizing C_R through *Principal Component Analysis*

$$C_R = R \Lambda R^T \, , \qquad t = -R \Lambda^{-1} R^T C_t.$$

- By considering an elliptical intersection (see Fig. 6), the canonical conic $C_{c1} x_{c1}^2 + C_{c2} x_{c2}^2 + C_{c3} = 0$ can be re-written as

$$\frac{x_{c1}^2}{a^2} + \frac{x_{c2}^2}{b^2} = 1 \quad \text{with} \quad a = \sqrt{-\frac{C_{c3}}{C_{c1}}}, \quad b = \sqrt{-\frac{C_{c3}}{C_{c2}}},$$

which can also be represented as a 2D Gaussian distribution $\mathcal{N}(\mu, \Sigma) = \mathcal{N}(t, R \, \Sigma_c \, R^T)$, where Σ_c is defined by

$$\Sigma_c = \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix}.$$

- By considering this distribution in our experiments, the likelihood $\mathcal{L}_{i,j}$ at time step t_j for an object i (located at initial position $x^{(i)}$) can then be defined by

$$\mathcal{L}_{i,j} = \frac{1}{\sqrt{(2\pi)^2 |\Sigma_j|}} \, e^{-\frac{1}{2}\left((x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right)}.$$

- When considering M different objects, a probabilistic measure of interest (level of saliency) for object i at each time step t_j is thus defined by (see also the bottom graph of Fig. 7)

$$p_{i,j}^G = \frac{\mathcal{L}_{i,j}}{\sum_{n=1}^{M} \mathcal{L}_{n,j}} \in [0, 1]. \tag{1}$$

TABLE III

DETECTION OF ATTENTIONAL UTTERANCES IN THE VOCAL TRACE.

OFFLINE TRAINING PHASE

- A set of 10 short common attentional utterances used as vocal spotlights (such as "Look here!" or "Watch this!") produced by the user is first recorded in a training phase prior to the interaction. A set of 10 random words and/or sentences spoken in a neutral way (e.g. by reading an instruction) is also collected.
- The pitch and energy of the sound signals are extracted, where the pitch (corresponding to the fundamental frequency f_0) is evaluated by the *subharmonic-to-harmonic ratio* method proposed by Sun [25].
- The two sets of pitch and energy traces are used to train two *Hidden Markov Models* (HMMs) λ^A and λ^N ("attentional model" and "neutral model"), where the number of states has been determined empirically. Each HMM is thus defined by 3 states, where each observation output is defined by a 2D Gaussian distribution (to encode the pitch and energy). The parameters are trained through the *Baum-Welch algorithm* [26], estimating iteratively the HMM parameters $\{\pi, A, \mu, \Sigma\}$, namely the initial state distribution π, the matrix of states transition probabilities A and the output distributions defined by centers μ and covariance matrices Σ.

ONLINE RECOGNITION PHASE

- When demonstrating a skill, the vocal trace of the user is recorded through the robot's internal microphone. It is then used to detect the probability of an attentional bid, i.e., when the user is bringing the attention of the robot to a particular aspect of the skill during the course of his/her demonstration.
- To do so, a temporal window of fixed size W is used to keep track of the pitch and energy signals during the last W seconds (if $t_j < W$, then $W = t_j$). We use here $W = 0.5$ sec., which has been determined empirically. The data in this window are then tested with the two HMMs λ^A and λ^N at each time step t_j through the *forward procedure* [26]. $\mathcal{L}_j^A$ and $\mathcal{L}_j^N$ are thus computed, corresponding to the likelihoods at time t_j of belonging respectively to the "attentional model" λ^A or to the "neutral model" λ^N.
- At each time step t_j, the probability of detecting an attentional utterance is finally defined by (see also fourth graph of Fig. 7)

$$p_j^S = \frac{\mathcal{L}_j^A}{\mathcal{L}_j^A + \mathcal{L}_j^N} \in [0, 1]. \tag{2}$$

most important, see e.g. [5], [27]. Another approach focuses on the prosody of the speech pattern rather than the exact content of the speech, which is used similarly to infer information on the user's communicative intent, see e.g. [14], [18], [28]. Some works also combine both information, see e.g. [23].

In this paper, we follow the prosodic approach by using *Hidden Markov Models* (HMMs) to detect particular intonation patterns while uttering attention to particular events during the demonstration of the task (e.g. emphasizing the use of a particular object).

Table III presents a method for the training and recognition of attentional cues using HMMs, showing how a vocal prior p^S can be retrieved from this model. The first four rows of Fig. 7 present an example of the result for the objects stacking task.

C. Combining several priors

We have seen in Section II that the statistical constraints relative to an object i are represented by the generalized trajectory $\hat{\mu}^{(i)}$ and associated covariance matrices $\hat{\Sigma}^{(i)}$. By

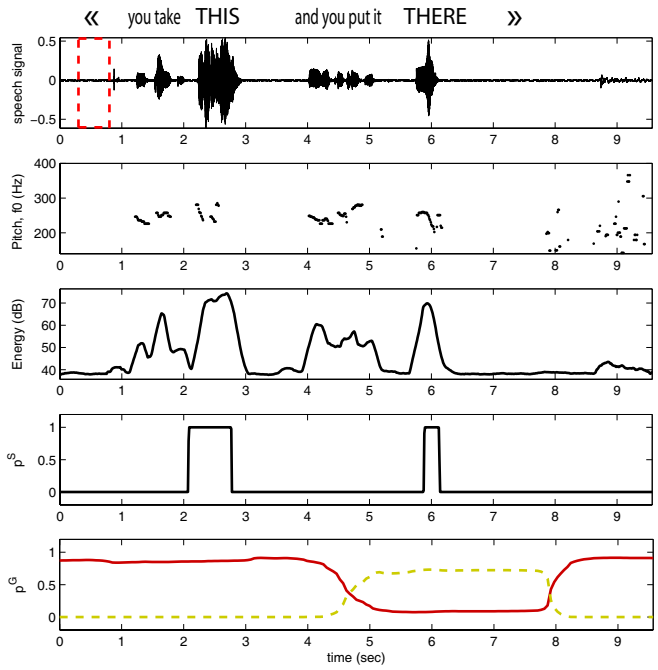

Fig. 7. Extraction of priors from speech (*first 4 rows*) and gaze information (*last row*) for the task depicted in Fig. 3. The first four graphs show the probabilistic extraction of attentional events in the vocal trace by using pitch and energy information (the temporal window of size W used to detect attentional cues is represented in dashed line). The first row shows the sound signal corresponding to the sentence *"You take THIS and you put it THERE"* told by the user when executing the skill (while observed by the robot, see left snapshot in Fig. 4). We see that the particular events in the demonstration, corresponding respectively to the subparts when the user grasps one object (*"THIS"*) and drop it on the other object (*"THERE"*), are highlighted through the user's voice. These events correspond roughly to local patterns characterized by a higher energy and a larger pitch amplitude with consecutive rising and falling intonation contours, which are typical to prosodic patterns serving as spotlights during the interaction [21], and which are automatically captured in our system through the HMM encoding. The third graph represents the probability p_j^S at time t_j of hearing an attentional utterance (see Table III). The bottom graph shows the probability p_j^G at time t_j of looking at the red cylinder (*in solid line*) or at the yellow cube (*in dashed line*), which also implicitly informs the robot that the user is conveying information on the relevance of these two objects at different time steps (see Table II).

using this GMR representation, we can modify easily the influence of the constraints by taking into consideration at each time step t_j gaze priors $p_{i,j}^G$ on object i and speech priors p_j^S. To do so, we first compute the mean values $\bar{p}_{i,j}^G$ and $\bar{p}_j^S$ at time step t_j by averaging over the different demonstrations provided to the robot (here, five). Then, we multiply by a weighting factor the covariance matrices of the GMR representation such that

$$\hat{\Sigma}_j^{(i)\prime} = \hat{\Sigma}_j^{(i)}(1 - \alpha\, \bar{p}_{i,j}^G\, \bar{p}_j^S), \qquad (3)$$

where $\bar{p}_{i,j}^G\, \bar{p}_j^S$ represents the joint probability at time step t_j (when considering speech and gaze as independent variables), which serves as a spotlight to emphasize particular events during the demonstration. α is a factor weighting the influence of the social cues over the constraints extracted through cross-situational statistics (here, $\alpha = 0.5$ has been selected empirically). Thus, for the reproduction of the task, the influence of the generalized trajectory with respect to object i is increased when $\bar{p}_{i,j}^G\, \bar{p}_j^S$ is high.

The imprecision due to the estimation of social cues is reduced by considering different demonstrations and different modalities. For example, in the bottom graph of Fig. 7, we see that from time step $t_j = 8$ sec., the system detects that the user is looking at the initial position of the cylinder (which is already stacked on the cube). This error may be due to tracking imprecision or because the user does not need to focus on a particular object/position anymore after he has dropped the cylinder. In both cases, this error disappears by taking into consideration the joint probability of events (i.e., the vocal analysis does not detect a particular event at these time steps), as well as the multiple demonstrations provided to the robot.

The right column of Fig. 5 shows the results of applying speech and gaze priors as proposed in Eq. (3) to the extraction of the task constraints (thinner envelope when $\bar{p}_{i,j}^G\, \bar{p}_j^S$ is high). By using social cues as priors, the robot can thus generalize the skill to different situations similarly to the reproduction results presented in Fig. 2.

IV. DISCUSSION AND FURTHER WORK

The early results presented above show that the integration of social cues within our statistical learning approach is promising. However, as only a very limited dataset has been used so far, the robustness of the approach still needs to be evaluated with untrained users teaching new skills in real-world experimental setups. The weighting mechanism defined in Eq. (3) to combine prosodic spotlights and joint attention spotlights is rather simple and serves as a first step towards evaluating the integration of social cues within the statistical framework. One direction of further work is to investigate the dependencies and relevance of these different cues in a human-robot teaching interaction context.

The advantage of using invasive devices such as the *X-Sens* motion sensors to track the user's gesture principally concerns the precision of the tracking procedure and robustness to occlusion. However, it is not an aim *per se* to use this modality, and a further step for user-friendly human-robot interaction will be to use the proposed framework with more human-equivalent modalities such as vision, see e.g. [1].

Ongoing work also concerns the joint use of this RPD framework with other methods such as a dynamical system in order to be robust to changes in the environment while the robot reproduces the learned skill [29], [30], or by using reinforcement learning as a way to let the robot explore its environment and learn by itself, thus extending the skill learned by imitation to a broader context than the one observed during the demonstrations [31].

Further work will extend the proposed scenarios to more complex interactions where the teaching phase and reproduction phase are more closely intertwined, allowing a richer interaction where the robot could request the user for advices at any time, where the user could provide advices on the robot's reproduction attempts, and where a more complex scaffolding process could be used. In this direction, one

short-term goal is to investigate how we could benefit from the probabilistic representation of the task constraints to segment the whole task into subtasks that can be reorganized differently to let the user provide scaffolds for each subpart independently at the desired speed and rhythm. Longer-term goals focus on developing robots that would have the capability to understand the user's intent from demonstrations, which would for example allow them to learn new skills even from failed attempts.

V. CONCLUSION

We presented a probabilistic approach in robot programming by demonstration that allows to extract incrementally the constraints of a task in a continuous form and to reproduce a generalization of the learned skill in new situations. We highlighted the importance of including the user's teaching abilities in the machine learning process, by using different modalities to convey the demonstrations (observational learning and kinesthetic teaching), and by designing human-robot interactive scenarios mimicking the human process of teaching. We then presented our current research towards a socially driven statistical learning framework to reduce the complexity of the skill transfer process. Through a manipulation task interaction with a humanoid robot, we illustrated how various social cues could be integrated in the proposed probabilistic framework to disambiguate automatically the role of the different variables/objects characterizing the task.

REFERENCES

[1] T. Spexard, M. Hanheide, and G. Sagerer, "Human-oriented interaction with an anthropomorphic robot," *IEEE Trans. on Robotics*, vol. 23, no. 5, pp. 852–862, 2007.

[2] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot programming by demonstration," in *Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Springer, 2008, in press.

[3] C. Nehaniv and K. Dautenhahn, "Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications," in *Interdisciplinary Approaches to Robot Learning*, J. Demiris and A. Birk, Eds. World Scientific Press, 2000, vol. 24, pp. 136–161.

[4] J. Saunders, C. Nehaniv, and K. Dautenhahn, "Teaching robots by moulding behavior and scaffolding the environment," in *Proc. ACM SIGCHI/SIGART Conf. on Human-Robot Interaction (HRI)*, March 2006, pp. 118–125.

[5] M. Pardowitz, R. Zoellner, S. Knoop, and R. Dillmann, "Incremental learning of tasks from user demonstrations, past experiences and vocal comments," *IEEE Trans. on Systems, Man and Cybernetics, Part B*, vol. 37, no. 2, pp. 322–332, 2007.

[6] T. Inamura, I. Toshima, and Y. Nakamura, "Acquiring motion elements for bidirectional computation of motion recognition and generation," in *Experimental Robotics VIII*, B. Siciliano and P. Dario, Eds. Springer-Verlag, 2003, vol. 5, pp. 372–381.

[7] S. Vijayakumar, A. D'souza, and S. Schaal, "Incremental online learning in high dimensions," *Neural Computation*, vol. 17, no. 12, pp. 2602–2634, 2005.

[8] S. Calinon, "Continuous extraction of task constraints in a robot programming by demonstration framework," PhD thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, 2007.

[9] Z. Ghahramani and M. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems*, J. D. Cowan, G. Tesauro, and J. Alspector, Eds., vol. 6. Morgan Kaufmann Publishers, Inc., 1994, pp. 120–127.

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

[11] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *Proc. ACM/IEEE Intl Conf. on Human-Robot Interaction (HRI)*, March 2007, pp. 255–262.

[12] ——, "What is the teacher's role in robot programming by demonstration? - Toward benchmarks for improved learning," *Interaction Studies*, vol. 8, no. 3, pp. 441–464, 2007.

[13] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[14] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo, "Tutelage and collaboration for humanoid robots," *Humanoid Robots*, vol. 1, no. 2, pp. 315–348, 2004.

[15] A. Thomaz and C. Breazeal, "Transparency and socially guided machine learning," in *Proc. IEEE Intl Conf. on Development and Learning (ICDL)*, 2006.

[16] Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto, "Responsive robot gaze to interaction partner," in *Proc. of Robotics: Science and Systems (RSS)*, August 2006.

[17] M. Riley, A. Ude, C. Atkeson, and G. Cheng, "Coaching: An approach to efficiently and intuitively create humanoid robot behaviors," in *Proc. IEEE-RAS Intl Conf. on Humanoid Robots (Humanoids)*, December 2006, pp. 567–574.

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

[19] P. Zukow-Goldring, "Caregivers and the education of the mirror system," in *Proc. Intl Conf. on Development and Learning (ICDL)*, 2004, pp. 96–103.

[20] G. Gergely and G. Csibra, "The social construction of the cultural mind: Imitative learning as a mechanism of human pedagogy," *Interaction Studies*, vol. 6, pp. 463–481, 2005.

[21] C. Yu and D. Ballard, "A unified model of early word learning: Integrating statistical and social cues," *Neurocomputing*, vol. 70, no. 13-15, 2007.

[22] M. Ito and J. Tani, "Joint attention between a humanoid robot and users in imitation game," in *Proc. Intl Conf. on Development and Learning (ICDL)*, 2004.

[23] G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. Christensen, "Situated dialogue and spatial organization: What, where... and why?" *Advanced Robotic Systems*, vol. 4, no. 1, pp. 125–138, 2007.

[24] S. Calinon and A. Billard, "Teaching a humanoid robot to recognize and reproduce social cues," in *Proc. IEEE Intl Symposium on Robot and Human Interactive Communication (RO-MAN)*, September 2006, pp. 346–351.

[25] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proc. IEEE Intl Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2002, pp. 333–336.

[26] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77:2, pp. 257–285, February 1989.

[27] P. Dominey, M. Alvarez, B. Gao, M. Jeambrun, A. Cheylus, A. Weitzenfeld, A. Martinez, and A. Medrano, "Robot command, interrogation and teaching via social interaction," in *Proc. IEEE-RAS Intl Conf. on Humanoid Robots (Humanoids)*, December 2005, pp. 475–480.

[28] F. Hegel, T. Spexard, B. Wrede, G. Horstmann, and T. Vogt, "Playing a different imitation game: Interaction with an empathic android robot," in *Proc. IEEE-RAS Intl Conf. on Humanoid Robots (Humanoids)*, December 2006, pp. 56–61.

[29] M. Hersch, F. Guenter, S. Calinon, and A. Billard, "Learning dynamical system modulation for constrained reaching tasks," in *Proc. IEEE-RAS Intl Conf. on Humanoid Robots (Humanoids)*, December 2006, pp. 444–449.

[30] E. Gribovskaya and A. Billard, "Combining dynamical systems control and programming by demonstration for teaching discrete bimanual coordination tasks to a humanoid robot," in *Proc. ACM/IEEE Intl Conf. on Human-Robot Interaction (HRI)*, 2008.

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