Hierarchical Multiagent Reinforcement Learning for Coordinating Verbal and Non-verbal Actions in Robots

Heriberto Cuayáhuitl[1] and **Nina Dethlefs**[2]

Abstract. This paper proposes an approach for learning to coordinate verbal and non-verbal behaviours in interactive robots. It is based on a hierarchy of multiagent reinforcement learners executing verbal and non-verbal actions in parallel. Our approach is evaluated in a conversational humanoid robot that learns to play Quiz games. First experimental results show evidence that the proposed multiagent approach can outperform hand-coded coordinated behaviours.

1 Introduction

Multiagent Reinforcement Learning is used to build autonomous agents that learn their behaviour from a shared environment [3]. In the case of cooperative Reinforcement Learning (RL) agents, they use the same reward function in order to optimize a joint goal [2, 12, 13, 10]. Recent research on interactive systems using machine learning has experienced important progress in the optimization of their conversational behaviours (e.g. confirmation, clarification and/or negotiation dialogues), where the RL framework has been an attractive alternative to hand-coded behaviours for the design of optimized dialogue agents. However, although important progress has been made for speech-based interactive systems, less progress has been made on optimizing both verbal and non-verbal behaviours in a unified way. Instead, both types of behaviours are often modelled independently [1, 15, 14, 8], without the aim to jointly achieve a goal as is the case in human interaction, where verbal and non-verbal behaviours are tightly coupled [16].

In this paper, we propose an approach based on hierarchical multiagent RL for optimizing the coordination of verbal and non-verbal behaviours. In this approach, one agent optimizes verbal behaviour, while another (simultaneously) optimizes non-verbal behaviour so as to align with the non-verbal actions of a human user. As a result, the joint action-selection of the RL agents represents the optimized coordination of both behaviours. We present preliminary results suggesting that this form of joint optimization is a promising and principled alternative to non-joint approaches and can equip robots with a more natural way of coordinating and adapting their multimodal actions.

2 Proposed Learning Approach

To achieve scalable dialogue optimization, we cast interaction control as a discrete-time Multiagent Semi-Markov Decision Process (MSMDP) $M = <S, \vec{A}, T, R, L, F>$ that is characterized by the following elements: (a) a finite set of states S; (b) a finite set of joint actions $\vec{A} = (A^v, A^{nv})$ executed in parallel, where A^v are verbal actions and A^{nv} are non-verbal actions; (c) a stochastic state transition function $T(s', \tau|s, \vec{a})$ that specifies the next state s' given the current state s and joint action $\vec{a} = (a^v, a^{nv})$, where τ denotes the number of time-steps taken to execute joint action $\vec{a}$ in state s; (d) a reward function $R(s', \tau|s, \vec{a})$ that specifies the reward given to the agent for choosing joint action $\vec{a}$ when the environment makes a transition from state s to state s'; (e) a language L that is represented as a context-free grammar (CFG) to represent relational tree-based representations as described in [4]; and (f) a stochastic model transition function $F = P(m'|m, s)$ that specifies the next model or subtask m' given model m and state s. This last element allows the user to navigate more flexibly across the available sub-dialogues [5].

We distinguish two types of actions: (i) single-step joint actions[3] corresponding to verbal actions such as 'greeting' or 'ask question' and non-verbal actions such as 'head nodding' or 'lift right arm', and (ii) multi-step joint actions corresponding to sub-dialogues or conjunctions of single-step joint verbal and non-verbal actions. In addition, we treat each multi-step joint action as a separate MSMDP.

We decompose an MSMDP into multiple MSMDPs that are hierarchically organised into X levels and Y models per level. The indices (i, j) only identify a unique subtask (i.e. MSMDP) in the hierarchy, they do not specify the execution sequence of subtasks which is learnt by the RL agent, where $j \in \{0, ..., X-1\}$ and $i \in \{0, ..., Y-1\}$. Thus, a given MSMDP in the hierarchy is denoted as $M^{(i,j)} = <S^{(i,j)}, \vec{A}^{(i,j)}, T^{(i,j)}, R^{(i,j)}, L^{(i,j)}, F^{(i,j)}>$. Notice that each MSMDP is a multi-decision maker for verbal and non-verbal actions, hence the term 'multiagent'. The solution to a Multiagent Semi-Markov Decision Process is an optimal policy $\pi^{*(i,j)}$, which is a mapping from environment states $s \in S$ to single- or multi-step joint actions $\vec{a} \in \vec{A}$. The goal of an MSMDP is to find a function denoted as $\pi^{*(i,j)}(s)$ that maximizes the cumulative reward of each visited state. The optimal policy for each learning agent in the hierarchy is defined by $\pi^{*(i,j)}(s) = \arg\max_{\vec{a} \in \vec{A}^{(i,j)}} Q^{*(i,j)}(s, \vec{a})$, where the optimal action-value function $Q^{*(i,j)}(s, \vec{a})$ specifies this cumulative reward for executing joint action $\vec{a}$ in state s and then following policy $\pi^{*(i,j)}$. We apply the HSMQ-Learning algorithm [9, 6] to cooperatively induce such a hierarchy of multiagent policies based on long-term cumulative rewards across policies.

3 Experimental Setting

To test our approach for generating coordinated joint actions and compare it with non-coordinated baselines, we use a robot dialogue

[1] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany, email: hecu01@dfki.de
[2] Heriot-Watt University, Edinburgh, Scotland, email: n.s.dethlefs@hw.ac.uk

[3] We assume that the execution of single-step joint actions terminates at the same time, which involves a non-verbal action to wait for the verbal one to terminate, or vice versa. Other ways of termination, where agents behave more autonomously but still in a coordinated way, are left as future work.

In Proceedings of the ECAI Workshop on Machine Learning for Interactive Systems: Bridging the Gap Between Language, Motor Control and Vision, Montpellier, France, pages 27-29, 2012.

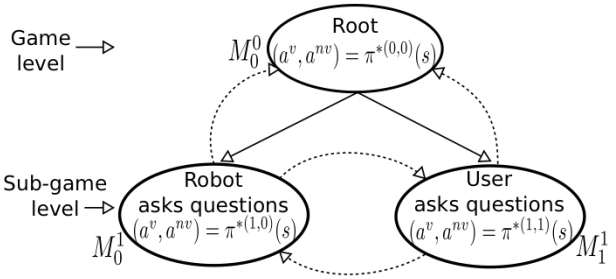


Figure 1. Hierarchy of joint agents for our robot in the Quiz domain. Whilst straight lines denote strict hierarchical control, dashed lines denote less strict control for more flexible interaction across sub-dialogues [5].

system playing Quiz games. In this domain, the robot can ask the user questions, or vice-versa, the user can ask the robot questions. Both user and robot communicate with verbal and non-verbal actions and our aim is to coordinate the robot's non-verbal actions with its verbal actions and simultaneously align them with the user's non-verbal actions to show individualised adaptation. Our system has been implemented using the Nao humanoid robot (see dialogue in Table 2).

We use the hierarchy of dialogue agents shown in Figure 1. Table 1 shows the set of state variables for our system, each one modelled as a discrete probability distribution with predefined parameters. Dialogue and game features are included to inform the agent of situations in the interaction. The set of verbal actions (80 in total) consists of meaningful combinations of speech act types and associated parameters.[4] The set of non-verbal actions (20 in total) consists of predefined body movements.[5] We constrained the actions per state based on the CFGs $L^{(i,j)}$, so that only a subset of joint actions was allowed per dialogue state (constraints omitted due to space). This reduces the state-action space from 10^{12}, using a propositional representation enumerating all variables and values, to only 10^4.

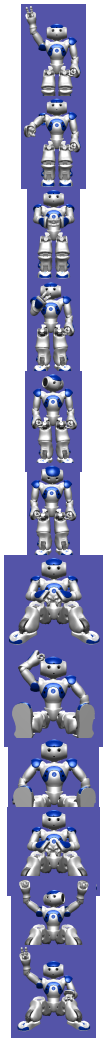
The global **reward function** aims for interactions that encourage to play, get as many correct answers as possible, and imitate the user's non-verbal actions. It is defined by the following rewards for choosing action a in state s: $+10$ for answering a question correctly or reaching a terminal state (in which the user will be prompted to play again), -10 for remaining in the same state (i.e. $s_{t+1} = s_t$ or $s_{t+1} = s_{t-1}$), $+1$ for imitating a non-verbal action, and 0 otherwise. The **multimodal user simulation** used a set of user dialogue acts as responses to the system dialogue acts (Footnotes 4-5). They used predefined probability distributions for modelling verbal and non-verbal interactions: $P(a^{v,usr}|a^{v,sys})$ and $P(a^{nv,usr}|a^{nv,sys})$, with errors based on an equally distributed speech and gesture recognition error rate of 20%. The recognition confidence scores were generated from beta probability distributions with parameters ($\alpha = 2$, $\beta = 6$) for bad recognition and ($\alpha = 6$, $\beta = 2$) for good recognition [4].

[4] **Verbal Single-Step Actions**: Speech Act Types={Salutation, Request, Apology, Confirm, Accept, SwitchRole, Acknowledgement, Provide, Stop, Feedback. Express, Classify, Retrieve, Provide.} × Parameters={Greeting, Closing, Name, PlayGame, Asker, KeepPlaying, GameFun, StopPlaying, Play, NoPlay, Fun, NoFun, GameInstructions, StartGame, Question, Answers, CorrectAnswer, IncorrectAnswer, GamePerformance, Answer, Success, Failure, GlobalGameScore, ContinuePlaying}

[5] **Non-Verbal Single-Step Actions**={Hello, Bye, HandShake, NodYes, NodNo, Success, Failure, OpenRightArm, OpenLeftArm, SitDown, StandUp, SeatedWithExtendedLegs, SeatedWithCrossedLegs, Thinking, ScratchingHead, StandingWithCrossedArms, StandingWithArmsBack, StandingWithArmsHeadBack, Wait, None.}

State Variable	Values
Salutation	none, greeting, withName, regreeting, closing
UserName	unknown, filled, known
ConfScore	null, 0.1, 0.2, 0.3, 0.4, 0.5, ... , 0.9, 1.0
Confirmed	null, no, yes
PlayGame	unknown, no, yes, ready
Instructions	unprovided, provided
Asker	unknown, robot, user
QuizGame	unplayed, playing, semiplayed, played, interrupted, keepPlaying, stopPlaying
GameFun	unknown, no, yes
GameOver	no, yes
GameInstructions	unprovided, provided
QuestionState	null, unknown, unasked, askedWithAnswers, askedWithoutAnswers, reaskedWithAnswers, reaskedWithoutAnswers, confirmed
AnswerState	unanswered, unclassified, correct, incorrect, unknown
MaxQuestions	no, yes
GameScore	unknown, good, bad
GlobalGameScore	null, unprovided, provided
ExpressedScore	no, yes

Table 1. State variables for the Quiz dialogue system, where combinations of variable-value pairs define situations (states) in the interaction used by the reinforcement learning dialogue agents for joint action-selection.



Hi, my name is NAO, what is your name?
I am Charles

Nice to meet you Charles.
Do you want to Play a Quiz game?
Sure

One of us asks questions and offers possible answers.
The other answers by selecting one of those answers.
Do you want to ask me questions or shall I ask you?
You ask me
The first question is. What is a typical magic instrument?
Your choices are: One, guitar. Two, drums.
Three, violin. Four, flute. What is your choice?
Violin

That is not quite right. Do you have another choice?
Maybe flute

Good, that is correct.
Now I want to ask you

Let's do it like that. What is the first question?
What does water become at the temperature of zero degrees?
Your choices are: One, condensation. Two, ice. Three, steam.

I choose ice.
That is correct.

Okay, ask me another question.
I want to stop playing.

Did you like playing the Quiz Game?
I did

I am glad to hear that.

It was nice playing with you, see you soon. Bye!

Table 2. Illustrative multimodal dialogue exhibiting non-verbal actions (left) and verbal actions (right). User responses shown in italics. The robot's images were generated with the Choregraphe tool from `aldebaran.com`

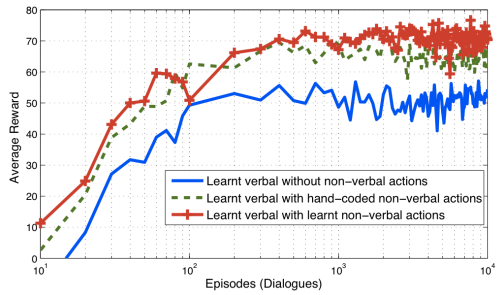


Figure 2. Average reward (10 runs) of joint action learners. Settings: $\alpha = 100/(100 + \tau)$, γ of .99, ϵ-Greedy, $\epsilon = .01$, initial Q-values$= 0.01$.

4 Experimental Results

We trained our agents, and compared their performance in terms of *dialogue reward* against two baselines; see Figure 2. One baseline uses *learnt verbal actions without non-verbal actions* (solid blue line), and the other baseline uses *learnt verbal actions with hand-coded non-verbal actions* (dashed green line). The latter baseline included intuitive joint actions such as *<Salutation(Greeting),Hello>* or *<Feedback(CorrectAnswer),NodYes>*. Results from the last 1000 episodes show that our multiagent approach (red crossed line) outperforms its counterparts (blue and green lines) by 27% and 8% in terms of average reward, respectively. We can draw the following preliminary conclusions. While the low performance of the verbal-only baseline most likely results from its lack of non-verbal expressiveness (and therefore lack of positive rewards for imitating the user), the difference between the jointly learnt and hand-coded policies is most likely related to adaptiveness. While the hand-coded policy relies on intuitive combinations of verbal and non-verbal actions, users differ with respect to their individually preferred combinations. Coordinating verbal and non-verbal actions jointly based on imitation of the user's gestures, therefore leads to a higher degree of individualised adaptation and higher rewards.

As a consequence of these results, we will investigate two hypotheses in future research: (1) a humanoid robot that only speaks but does not move has a lower perceived performance than a robot that combines verbal with non-verbal actions; and (2) a humanoid robot that does not learn to coordinate its verbal with non-verbal actions in an adaptive fashion is perceived as having a lower performance than a robot that learns to coordinate both types of actions. An advantage of learning to coordinate verbal with non-verbal actions is that the robot can exhibit different behaviours for different users. Future work may also investigate how coordinated verbal and non-verbal behaviour may affect task success or user satisfaction.

5 Conclusion and Future Work

We have described an approach for optimizing the behaviour of robot dialogue systems by applying and extending a hierarchical RL framework to support multiagent decision making of verbal and non-verbal actions in a coordinated and adaptive way. To evaluate, we have incorporated our methods into a robot dialogue system that learns to play Quiz games. Although preliminary, experimental results make our approach look promising by combining the benefits of (a) predefined state-action spaces, (b) scalable policy learning, (c) joint and coordinated action section, and (d) opportunities for online learning.

We argue that those features, with a special focus on online learning, represent an interesting direction to train robots' behaviour, so that they can learn how to coordinate their actions in an adaptive fashion while interacting with users. The next step towards this is to train our simulations and MSMDPs (online) from real human-robot interactions to validate our results. We would like to optimize turn-taking for more natural and efficient interactions. Another step is a comparison with other hierarchical learning algorithms [11] using function approximation. We also would like to extend our joint learning agents with adaptive verbalizations [7], where each MSMDP in our hierarchy of agents would have three agents, one for dialogue management, one for language generation, and one for non-verbal behaviour.

6 Acknowledgments

This research was funded by the European FP7 programmes under grant agreements ICT-248116 (ALIZ-E) and 287615 (PARLANCE).

REFERENCES

[1] Dan Bohus and Eric Horvitz, 'Facilitating Multiparty Dialog with Gaze, Gesture, and Speech', in *ICMI-MLMI*, p. 5, (2010).
[2] Craig Boutilier, 'Sequential Optimality and Coordination in Multiagent Systems', in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 478–485, (1999).
[3] L. Busoniu, R. Babuska, and B. De Schutter, 'A Comprehensive Survey of Multiagent Reinforcement Learning', *IEEE Transactions on Systems, Man, and Cybernetics*.
[4] H. Cuayáhuitl, 'Learning Dialogue Agents with Bayesian Relational State Representations', in *IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Barcelona*, pp. 9–15, (Jul 2011).
[5] H. Cuayáhuitl and I. Kruijff-Korbayová, 'An Interactive Humanoid Robot Exhibiting Flexible Sub-Dialogues.', in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Montreal, Canada*, (Jun 2012).
[6] H. Cuayáhuitl, S. Renals, O. Lemon, and H. Shimodaira, 'Evaluation of a Hierarchical Reinforcement Learning Spoken Dialogue System', *Computer Speech and Language*, **24**(2), 395–429, (2010).
[7] N. Dethlefs and H. Cuayáhuitl, 'Hierarchical Reinforcement Learning for Adaptive Text Generation', in *International Conference on Natural Language Generation (INLG)*, Dublin, Ireland, (Jul 2010).
[8] N. Dethlefs, V. Rieser, H. Hastie, and O. Lemon, 'Towards Optimising Modality Allocation for Multimodal Output Generation in Incremental Dialogue', in *ECAI Workshop on Machine Learning for Interactive Systems: Bridging the Gap Between Language, Motor Control and Vision (ECAI-MLIS), Montpellier, France*, (Aug 2012).
[9] T. Dietterich, 'An Overview of MAXQ Hierarchical Reinforcement Learning', in *Symposium on Abstraction, Reformulation, and Approximation (SARA)*, pp. 26–44, (Jul 2000).
[10] M. Ghavamzadeh and S. Mahadevan, 'Hierarchical Multiagent Reinforcement Learning', *Journal of Autonomous Agents and Multi-Agent Systems*, **13**(2), 197–229, (2006).
[11] Bernhard Hengst, 'Hierarchical Reinforcement Learning', in *Encyclopedia of Machine Learning*, 495–502, (2010).
[12] M. Lauer and M. Riedmiller, 'An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-agent Systems', in *International Conference on Machine Learning (ICML)*, pp. 535–542, (2000).
[13] M. Lauer and M. Riedmiller, 'Reinforcement Learning for Stochastic Cooperative Multi-Agent-Systems', in *Intl. Confrence on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 1516–1517, (2004).
[14] V. Rieser and O. Lemon, 'Learning and Evaluation of Dialogue Strategies for New Applications: Empirical Methods for Optimization from Small Data Sets', *Computational Linguistics*, **37**(1), 153–196, (2011).
[15] A. C. Tenorio-Gonzalez, E. F. Morales, and L. Villaseor-Pineda, 'Dynamic Reward Shaping: Training a Robot by Voice', in *Ibero-American Conference on AI (IBERAMIA), Bahía Blanca, Argentina*, (Nov 2010).
[16] A. Yamazaki, K. Yamazaki, M. Burdelski, Y. Kuno, and M. Fukushima, 'Coordination of Verbal and Non-verbal Actions in Human-Robot Interaction at Museums and Exhibitions', *Journal of Pragmatics*, **42**(9), 2398–2414, (2010).