Learning socio-communicative behaviors of a humanoid robot by demonstration

Duc-Canh Nguyen

UNIVERSITÉ GRENOBLE ALPES

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Signal Image Parole Télécoms**

Arrêté ministériel : 7 août 2006

Présentée par
Duc-Canh NGUYEN

Thèse dirigée par **Gérard BAILLY** et
codirigée par **Frédéric ELISEI**

préparée au sein du
GIPSA-lab
dans **l'École doctorale Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)**

Learning socio-communicative behaviors of a humanoid robot by demonstration

Thèse soutenue publiquement le **22 octobre 2018**,
devant le jury composé de:

Mohamed CHETOUANI
Professeur, UPMC, ISIR/Paris (Rapporteur)
Philippe GAUSSIER
Professeur,UCP, ETIS/Cergy Pontoise (Rapporteur)
Denis PELLERIN
Professeur,UGA, GIPSA-lab/Grenoble, (Examinateur)
Mathieu LEFORT
Maître de conférence, UCB, LIRIS/Lyon (Examinateur)
Gérard BAILLY
Directeur de recherche CNRS, GIPSA-lab/Grenoble (Directeur de thèse)
Frédéric ELISEI
Ingénieur de recherche CNRS, GIPSA-lab/Grenoble (Co-directeur de thèse)

Acknowledgements

First and foremost I would like to thank my advisors, Gérard Bailly and Frédéric Elisei, who gave me the opportunity to do this thesis. Over the course of three years, they have constantly offered their guidence and support, and spent time and energy in helping me to dive into new and complementary areas of research.

I would also like to thank my two rapporteurs (Mohamed Chetouani and Philippe Gaussier) and other jury members (Denis Pellerin and Mathieu Lefort), who helped me to improve this thesis.

A very special gratitude Agence nationale de la rechercher (ANR) and SOMBRERO project designed by Gipsa-lab for helping and providing the funding for the work.

I am grateful to my litter family, espcially my wife Van Anh and my daughter Myla , who have provided me through moral and emotional support in my life. I am also grateful to my other family members and friends who have supported me along the way.

And finally, last but by no means least, also to everyone in the Gipsa-lab, it was great sharing laboratory with all of you during last three years.

Contents

Introduction **1**

1 Social Robots **5**

1.1 Potential of Assistive Robots in our Society 5

1.2 Defining social Robots 6

1.2.1 Definition & Classification 7

1.2.2 Humanoid robots 7

1.2.3 Why Social Robots as Assistants? 9

1.2.4 Socially Assistive Robot (SAR): Long-Term vs. Short-Term Interactions 11

1.2.4.1 Long-term Interactive Robot 11

1.2.4.2 Short-term Interactive Robots 13

1.2.5 Our work 15

1.3 Learning Methods 16

1.3.1 Developmental Robotics 16

1.3.2 Learning by Demonstration 17

1.3.2.1 Demonstrating Low-level skills 17

1.3.2.2 Learning high-level behaviors - multimodal interactive behaviors 18

1.3.2.3 Incremental vs. Batch Learning 18

1.4 The SOMBRERO Framework 19

1.5 Summary 21

2 Human-Human Interactive Data: experimental design, acquisition, annotation & characterization **23**

2.1 "Put That There" (PTT) data 25

2.1.1 Scenario 25

2.1.2 Data Acquisition . . . 26
2.1.3 Data Annotation . . . 27
2.1.4 Comments . . . 29
2.2 Selective Reminding Test data (RL/RI) . . . 30
2.2.1 Scenario . . . 30
2.2.2 Data Acquisition . . . 31
2.2.3 Annotation . . . 32
2.2.4 Comments . . . 34
2.3 Conclusion and Discussion . . . 34

3 Multimodal Interactive Behavioral Models 37

3.1 State of the art: modeling multimodal interactive behaviors . . . 39
3.1.1 Rule-based methods . . . 39
3.1.2 Machine learning methods . . . 43
3.2 Recurrent Neural Network - Long-Short Term Memory . . . 43
3.2.1 Recurrent neural networks . . . 44
3.2.1.1 Simple Recurrent Neural Network . . . 44
3.2.1.2 Long-Short Term Memory . . . 45
3.2.2 Application of RNNs in human interactions . . . 46
3.3 Generating discrete events: Arm and Gaze . . . 48
3.3.1 Modeling techniques . . . 49
3.3.1.1 Hidden Markov Models . . . 49
3.3.1.2 Dynamic Bayesian networks . . . 51
3.3.1.3 Long-Short Term Memory . . . 52
3.3.2 Results and Discussion . . . 53
3.3.2.1 Off-line task . . . 53
Prediction accuracy . . . 53

Coordinate Histogram 54
3.3.3 On-line tasks 54
3.3.4 Discussion 57
3.4 Generating discrete variables from continous estimations: Backchannels 60
3.4.1 State of the art 60
Rule-based 60
Data-driven 60
3.4.2 Interactive data: Train and validation data 61
3.4.3 Methodologies 61
3.4.3.1 Conditional Random Field 62
3.4.3.2 Input window 63
3.4.3.3 Implementation 63
3.4.4 Backchannel prediction and generation 63
3.4.5 Subjective Evaluation 65
3.4.5.1 Lexical Choice 66
3.4.5.2 Relevance of BC generation 66
3.4.6 Discussion 67
3.4.7 Conclusion 68
3.5 Generating continuous variables: Head Motions 69
3.5.1 State of the art 69
3.5.1.1 Head motion, gaze and speech 69
3.5.1.2 Generating head motions 70
3.5.2 Multimodal interactive behavioral models for generating head motions . 71
3.5.2.1 Analyzing data 72
3.5.2.2 Results 74
3.5.2.3 Discussion 76

3.5.3 Conclusions & perspectives . . . 77
3.6 Summary . . . 79

4 Gesture Controllers: Design and evaluation 81

4.1 Adapt the *RL/RI* scenario from HHI to HRI . . . 83
4.1.1 Substituting sheets of paper with displays . . . 83
4.1.2 Dealing with response times . . . 84
4.2 Designing gesture controllers . . . 85
4.2.1 Speech . . . 85
4.2.2 Arm gestures . . . 87
4.2.3 Gaze . . . 87
4.2.4 Eyelids . . . 88
4.3 Evaluating gesture controllers . . . 89
4.3.1 Evaluation of HRI systems: state of the art . . . 89
4.3.2 Designing and performing on-line vs off-line evaluation . . . 90
4.3.2.1 The first evaluation . . . 92
4.3.2.2 The second evaluation . . . 94
4.3.3 Comparing the two evaluations . . . 95
4.3.3.1 Yuck responses . . . 95
4.3.3.2 Subjective ratings . . . 98
4.3.3.3 Comments . . . 98
4.4 Conclusions . . . 99

5 Towards Autonomous Robots and Evaluations 101

5.1 Towards an autonomous robot performing the *Put That There* scenario . . . 102
5.1.1 Required Modules . . . 102
5.1.2 A strategy to assess/evaluate the modules . . . 103

5.1.3 The robot replicating the *PTT* scenario . . . 105
5.1.4 Comments . . . 107
5.2 A basic (immature) Autonomous Robot performing the *RL/RI* scenario . . . 107
5.2.1 Rules-based interactive models for the *RL/RI* scenario . . . 107
5.2.2 An Autonomous Control Framework . . . 109
5.2.3 Comments . . . 111
5.3 Challenges remaining to achieve a widely acceptable autonomous robot . . . 112
5.4 Benefits of the Wizard of Oz approach . . . 114
5.4.1 The Beaming System . . . 115
5.4.2 Enhancing the system for the *RL/RI* scenario . . . 116
5.4.3 Evaluating the Immersive Teleoperation system . . . 117
5.4.3.1 An Evaluation Framework . . . 119
5.4.3.2 Item selection . . . 119
5.4.3.3 Illustrative Examples of Evaluating Data . . . 119
5.4.4 Comments . . . 121
5.5 Conclusion . . . 122

6 Conclusions and Perspectives 125

6.1 Conclusions . . . 125
6.2 Perspectives . . . 126
6.2.1 Evaluation Framework . . . 127
6.2.2 Collecting and Annotating Interactive Data . . . 127
6.2.3 Improving and Adapting Robots' Behaviors . . . 127
6.2.3.1 Sharing Control . . . 127
6.2.3.2 Generating variably interactive behaviors . . . 128
6.2.3.3 Adapting the Interactive Models: Transfer Learning . . . 128

A Sensorimotor Calibration for Pointing **131**

A.1 Comments 132

B Finite State Machine of the RL/RI scenario **135**

B.1 FSM 135

B.2 FSM models of the sub-tasks of the RL/RI scenario 136

B.3 Comments 138

List of publications **141**

Resumé **161**

Abstract **161**

List of Figures

1.1 Classifying social robots according to function and appearance [HNI] 8

1.2 Comparing robot and display in guiding and recommending shopping [Kan+09] 9

1.3 Long-term interaction robot: companion robots 12

1.4 Robovie Robots 13

1.5 KiliRo, a parot robot, interacting with autism children [Bha+17]. 14

1.6 Examples of short-term robot interaction 15

1.7 The robot is chatting with custom, recommending them shopings/ restaurants and using deictic gestures to guiding them in a mall [Kan+09]. 15

1.8 *Nina*, the iCub2 humanoid robot with mouth and lips articulation which has been used in this work. 19

1.9 The three main steps of learning interaction by demonstration. 20

2.1 Collect representative interactive behaviors from human coaches in HHI scenario 24

2.2 Table of "Put That There" scenario. 25

2.3 First-person view of the interaction captured from the instructor's head-mounted scene camera. 26

2.4 The instructor's head and right arm movements are monitored by the MoCap system. 27

2.5 Semi-automatic segmentation of arm movements according to speech and target of the pointing gesture. 29

2.6 Speech production waiting for arm movement planning 30

2.7 Capturing the multimodal behavior of the human tutor during HHI. 31

2.8 Semi-automatically annotated data with Elan software. 33

2.9 Number of occurrences of the 34 different lexical markers used by the interviewer to encourage the subjects [Bai+16]. 33

2.10 Timing of ends (blue) and beginnings (rose) of interlocutors' speeches surrounding backchannels. 34

3.1 Face-to-face interaction within the time scale of human actions [Thó99] 38

3.2 SOMBRERO learning framework: Modeling multimodal interactive behavioral models 39

3.3 The three stages of SAIBA and the two mediating languages: FML (function markup language) and BML (behaviour markup language) 40

3.4 An example of a BML block [Kop+06] 40

3.5 Situated modules controlled by episode rules [Kan+02] 42

3.6 A multimodal interactive behavioral modal using DBN for a humanoid robot in a narration task [HM14]). 44

3.7 Unfold RNN 44

3.8 RNN vanishing problem [Den] 46

3.9 The LSTM unit is a memory block that can be updated, erased or read out according its internal activation c_t and the current input x_t. 47

3.10 LSTM unroll. 47

3.11 Bi-directional RNN 48

3.12 Schematic of HMM-based multimodal interactive modeling 50

3.13 The learned structure of the DBN model 51

3.14 Schematic model of multi-tasking LSTMs 53

3.15 Offline generation: comparing performance of the joint estimation 55

3.16 Input and output sequences 56

3.17 Computing coordinate histogram corresponding to SP by cumulating delays between each SP event and adjacent events in the other two streams GT and FX. 57

3.18 Comparing ground truth coordination histograms with those computed with streams predicted by different offline methods 58

3.19 Performance of the different methods for the on-line prediction tasks. Same conventions as for Figure 3.15 59

3.20 Inputs concatenated w-context windows 63

3.21 Precision-Recall curve of backchannel prediction with CRF and LSTM 64

3.22 Backchannel prediction and generation by CRF vs. LSTM from speech activities of both speakers. 65

3.23 Number of backchannels generated by different methods 66

3.24 Subjective evaluation results 67

3.25 Inputs with concatenated two frames t and $t+w$ 68

3.26 Correlation of CCAs between each of H1, H2, H3 and FX, IU, GT, SP, MP and F0 72

3.27 Single vs. Multi-task models generating head motions 73

3.28 Average H1 RMSE at different epochs corresponding to the different cascaded models. 74

3.29 (a) H1 real vs. prediction streams between different models; (b) input streams (FX ground truth& SP) and FX prediction from LSTM1. 75

3.30 Coordination histograms among H1 and (IU,SP). 76

3.31 Chi-squared distances of the different prediction models. 77

3.32 Average H1 RMSE of the *Baseline* model without and with SP shifted frames corresponding to number of training epoch. 78

3.33 CCA of Hs vs. SP with various number of shifted frame. 79

3.34 Average H1 RMSE without and with SP shifted frames corresponding to the different models. 80

4.1 Gesture controllers: design and evaluation. 82

4.2 Adapted RL/RI scenario for human-robot interaction 83

4.3 An example of different durations between a robot action and a human action when performing the same event. 84

4.4 Predicting prosody with superposition of functional contours 86

4.5 Robot's arm. 87

4.6 Two examples of robot's gaze. 88

4.7 Robot's eyelids. 88

4.8 Schematic model of continuous quality rating with scale [HK99]. 90

4.9 The *Nina* robot from the subject's perspective. 91

4.10 General framework of evaluation gesture controllers. 92

4.11 The yucking probability as a function of time for first by participants. 93

4.12 Density of yuck responses for our replayed interaction. 93

4.13 Comparing the yucking probability as a function of time for first vs. second assessment by the subjects. 96

4.14 Comparing subjective ratings according to conditions (same conventions as figure 4.13). 96

4.15 Overall repartition of ratings to questions 1 and 3 according to sex 97

5.1 A strategy of evaluating the robot step-by-step: from a replicated version to an autonomous version. 104

5.2 Excerpts from the PTT replicating scenario. 106

5.3 Schematic model of a baseline autonoumous robot interacting with a human subject in the RL/RI scenario. 109

5.4 Autonomous robot performing RL/RI task with human subject using rule-based model 111

5.5 Collecting Human-Robot Interaction data with immersive teleoperation 113

5.6 Immersive teleoperation system used to collect interactive data for the *RL/RI* . 116

5.7 An example of the robot using the Beaming system performs the *RL/RI* scenario.117

5.8 Beaming evaluation using a virtual tablet (c) and a screen for items (b) 120

5.9 Distributed distance between two consecutive items 121

5.10 The items are sorted and merged to exhibit a balanced order 121

5.11 Gaze and head movement of a subject in clicking-beaming evaluation 122

A.1 Different possible approaches for mapping pointing gestures [Lem+13]. 132

A.2 Schematic model of arm pointing with laser spotter attached on the index finger.132

B.1 A simple FSM model of a Turnstile 135

B.2 A more detailed FSM model of a Turnstile 136

B.3 A sub-FSM modeling the learning phase 137

B.4 Sub-FSMs of counting and testing phase 138

B.5 A sub-FSM of recognition phase 139

List of Tables

2.1 Semi-active labeling discrete events 32

3.1 Chi squared distances between the coordination histograms of ground truth vs. those of the different off-line models 54

3.2 Chi-squared distances between the coordination histograms of ground truth vs. those of the different on-line models 57

3.3 Precision, recall and F1 score of the two methods in BC prediction task. 64

3.4 F1 score of the two methods with inputs concatenating current frame and only a past frame with distance w. 68

3.5 Head motions generated by rules [LM06] 71

3.6 Root mean square errors (Pearson correlations) between ground truth and predicted head motions with different models. 75

4.1 Adapting events from HHI to HRI 85

4.2 Causes of yuck behaviors of the first evaluation and modifications for the second evaluation 94

Introduction

Recently, socially assistive robotic (SAR) that not only helps people via physical interactions but also via social skills, have been receiving attention by researchers. This is due to the fact that SARs can engage in conversation with humans or even create strong affective relations with humans. One major goal of building SARs is to build robots that are accepted by people in human-centered environments. Because humans interact with each other through complicated bidirectional multimodal signals to convey and perceive information, SARs need to be designed to conform to human expectations in communication. The robot behaviors have to follow human communication norms, social values and standards so that people might be able to intuitively understand the robots just as they could understand other people [Bre04]. Otherwise, people will be confused and have difficulties to anticipate reactions, even reject robot interactions. Therefore, SARs need to be provided with multimodal interactive behaviors, which are fluently coordinated with verbal (speech) and non-verbal (e.g. arm gestures, eye movements) behaviors of its human partners.

Depending on the objectives of Human-Robot Interaction (HRI), SAR faces different challenges. One of the important dimensions is the duration of this interaction. Some of SARs are concerned with long-term (LT) interaction, aiming at providing users with social glue and affective relations. Others are designed to engage into more short-term (ST) task-oriented interactions with a large range variety of users who are mostly unknown beforehand. The challenge of ST interactions is more oriented towards attention and quick adaptation.

The objective of this thesis is to endow a humanoid robot with sociocommunicative abilities in order to perform ST taks-oriented interactions. A way to provide the robot with ST interaction skills is to teach the robot behaviors by demonstration. In this thesis, our robot learns social skills from a human coach by a series of demonstrations of situated Human-Human interactions (HHI). This HHI-based framework faces several challenges:

Adaptability to different humans and situations Social robots should adapt their behaviors to interact with humans with different social profiles (personality, gender, age, emotional state, etc.). For example, introvert people prefer to interact with a robot that expresses introvert behaviors **thirobot**(e.g. speaking slowly, moving its arms narrowly, praising users). In contrast, extrovert people are more likely to prefer extrovert behaviors of robots (e.g. speaking fast, opening its arms widely). robots should also adapt their language level (lexicon, syntax ...) with regards to the linguistic competence of the human partner, such as using simple phrasal constructions with children.

Transferring skills from HHI to HRI Another problem is to scale the human model to the specific interaction capabilities of the robot in term of physical limitations (degrees of freedom), perception, action and reasoning. For example, robots have imperfect perception modules (e.g. speech recognition, especially to recognize infants' or elderly speech). Or the robot could not have a flexible hand to open, close or write on a notebook

using a pen as humans do. Therefore, when transferring HHI-based data, the scenario or the interaction tools need to be re-designed to cope with the current robot sensorimotor abilities.

Modeling of joint interactive behaviors Because humans tend to apply social models when interacting with robots [BSS11], SAR need to be endowed with multimodal interactive models. The modeling of multimodal interactive behaviors – that consists in mapping the partenes' sensorimotor streams – is rather challenging since human-human interaction are paced by complex perception-action loops.

To generate natural behaviors, the models have to capture intra- vs inter-coordination between the modalities. Intra-coordination means relationships between modalities inside the robot (e.g. robot speech, gaze, head motions and hand pointing gestures). In contrast, inter-coordination relates to the coordination between modalities of the robot and its human partners (e.g. turn taking or backchanneling).

The replay and evaluation of these behaviors by the robot In order to train and evaluate the models, objective metrics are often used in machine learning methods such as maximizing F1-score or minimizing RMS errors. However, in the context of human-robot interaction, this should be completed with subjective evaluation, in order to measure the satisfaction (succesful and meaningful actions) of the human users with regards to the robot's behaviors. In this thesis, we propose a method for the online evaluation of interactive behaviors that provide timestamps of the robot's faulty behaviors.

The SOMBRERO project designed in our laboratory aims to solve all of these challenges in order to provide an iCub humanoid robot with social skills for ST interactions. My thesis contributes to the project by addressing the two last challenges: (3rd) building interactive multimodal interactive models to generate adequate robot actions; (4th) implementing and evaluating the actions on the robot. Also, we give discussions and our perspectives to solve the two first challenges in the context of the SOMBRERO project.

With regards to the third challenge, there are two main approaches to model multimodal interactive behaviors: rule-based vs. machine-learning methods. The rule-based methods with hand crafted rules are time consuming and require a lot of human labor. In addition, they have difficulties in taking into account many factors conditioning the multimodal behaviors (task, personality, social context, emotion, gender, etc.). In contrast, the machine-learning methods are expected to solve this problem by automatically finding behavioral regularities from data. Up to now, few machine-learning methods have been applied to interactive multimodal data for building behavioral models. Most of them are often based on statistical methods such as Hidden Markov Model, Dynamic Bayesian Networks, etc. or more sophisticated graphical models capturing causal relations between multimodal data. However, these methods have difficulty in capturing long-term temporal dependencies between random variables, which are crucial for the coordination of modalities. In this work, we explore deep learning (DL) with recurrent connections to deal with the time dependencies. However, training end-to-end DL models usually requires a lot of interaction data which are not always available due to expensive

collecting cost. We however show that DL models can outperform other statistical models in generating the robot's behaviors, even with a limited training dataset.

The fourth challenge faces two issues. Firstly, we have to design gesture controllers to drive robot actions. These controllers should exploit the robot's sensorimotor capabilities but also be judged equivalent to the social behaviors of human coaches. For example, a human-like hand pointing gesture should move fast towards the target. This control policy is not optimal for a robot that would prefer to slow down progressively to minimize vibrations and optimize accuracy. A robot without eyebrows nor frowning capabilities would have difficulties in signaling astonishment or irritation. This should be compensated with enhanced verbal or other coverbal behaviors. Secondly, evaluating robot behaviors requires a method to detect WHEN and WHAT wrong social behaviors occur so that these behaviors can be repaired in order to increase human acceptance.

This thesis manuscript is organized in five chapters:

In the first chapter, we will briefly introduce SARs, motivation of SARs, long-term (LT) vs. short-term (ST) applications of SARs. We also introduce approaches to learn robot's behaviors. We introduce the SOMBRERO project, in which this work is inscribed. this project proposes a learning framework to solve the two first challenges for HHI-based frameworks: scale human-human interaction to human-robot interaction and compensate for possible drastic changes of human behaviors in front of robot by using demonstrations from immersive teleoperation of the robot.

In the second chapter, we detail our ST interactive scenarios and how we process and annotate the data in order to get useful features to train interactive models. The robot will be involved in two scenarios: (a) *a neuro-psychological test* (RL/RI) in which the robot intends to play the role of a psychologist interacting with elderly people and performing a memory test; (2) a collaborative task named *Put That There* (PTT) in which the robot instructs a manipulator how to move cubes. Both tasks require the fluent coordination of co-verbal and nonverbal behaviors such as gaze, head, arm movement etc. with adequate verbal behaviors. We explain how we convert raw signals into discrete events before building interactive models.

In the third chapter, we develop and train multimodal interactive behavioral models that generate appropriate action events from perception streams. The models could then be used to generate coverbal behaviors for our iCub humanoid robot. We here focus on a Deep Learning architecture named Long-Short Term Memory (LSTM) which can capture temporal dependencies between social signals over a long period of time. We compare its performances with several statistical models such as Hidden Markov Model (HMM), Dynamic Bayesian Network (DBN), Conditional Random Field (CRF). We show that, due to the ability of capturing long-term dependencies between hidden states, the best accuracy of generating gaze and arm movements is achieved by the LSTM method.

We also used the LSTM method to generate continuous head motions. Using Canonical Correlation Analysis (CCA), we found that the gaze highly correlates with head motion

in the PTT corpus. We propose a *cascaded* LSTM model: a first LSTM layer is used to predict gaze and its output is further fed into another LSTM layer generates head motions. The two LSTM models can be trained separately and the cascaded model further fine-tuned. This cascaded model – that explicitly takes into account the causal relations between saccades and head movements – performs a better prediction of head motion than a baseline predicting head motion & gaze from speech activities without any a priori hierarchy. In this chapter, we also illustrate how LSTM and Conditional Random Field (CRF) can be used to generate verbal back-channels from speech activities and how to use window contexts to improve prediction and generation results.

The fourth chapter mainly focuses on the 4th challenge. This chapter proposes to adapt the RL/RI scenario in order to cope with the limitations of the robot's motor cabilities. We design and evaluate gesture controllers to execute actions events (generated by multimodal interactive behavioral models). We also propose an original on-line evaluation framework that not only detects the robot's faulty behaviors but also suggest how to iteratively improve the quality of the robot's behaviors so that the robot can achieve higher acceptance rate by humans.

The fifth chapter presents our work in progress to build autonomous robots that can perform interactive scenarios as well as how to improve social skills for the autonomous robots in the near future. For the *Put That There* scenario, robot can autonomously perform gaze and pointing gestures with the trained interactive models. Further, we give some perspectives on building perception modules which have been not developed on this thesis. For the *RL/RI* scenario, we design a first autonomous robot with gesture controllers and backchannels generated by interactive models described in 3th chapter and some additional rules. To improve the robot behaviors, we propose to establish experiments to collect social signals for the robot using immersive teleoperation. We also depict our plan to evaluate and improve our current beaming system.

Finally, we give **conclusions and perspectives** about our approaches and future works about how to improve quality of robot behaviors and complete the autonomous control of our humanoid robot.

Chapter 1

Social Robots

We have found that individuals' interaction with computers, television and new media are fundamentally social and natural, just like interactions in real life. Reeves et al. [RN96]

This chapter motivate applications of the Socially Assistive Robots (SAR). The chapter also introduces the state of the art of methodologies used to teach such robots how to perform specific tasks. We finally sketch our approach to build a SAR within the SOMBRERO framework.

1.1 Potential of Assistive Robots in our Society

There is a dramatic increase of elderly population. Today, the world's population over 60 is approximated about 10 percents and the number can double by 2050 [Pol05]. Especially, there are 10 percent of elderly persons over 65 and 50 percent over 85 with Alzheimer's disease [Heb+03]. One of the specific of sensory-motor and psychosocial issue of aging is cognitive decline. Therefore, there will be fewer and fewer young people for assisting the elderly people to cope with the challenges of aging. Also, the cost of nursing home for elderly may increase dramatically in the future [Roy+00]. In addition, people prefer to live in their own homes as long as possible rather than in nursing houses when they have health problem concerning age. One of the major challenges of our society is to take care of elderly people personally at home at low costs.

Recently, robotics technologies are growing with a major revolution. With a considerable cheaper computational cost and increasing significant quality as well as quantity of sensor technologies (e.g. speech recognition, machine vision), and remarkable improvement of machine learning in particular thanks to the performances reached by deep learning techniques (e.g. image recognition, speech generation, etc.), we are closer to the goal of intelligent service robots that can assist people in their daily living activities more than ever before. Therefore, robots will become more and more popular and be able to interact with people in human environments in near future.

These service robots promise to solve the growing personnel shortage, which are used in daily life for taking care of the elderly people, motivating cognitive and physical exercise. For

example, robots might assist people in cooking, cleaning, talking or provide treatments in emergency stroke, involving people with cognitive tasks such as reminding medication schedules, maintaining healthy habits as well as socially engaging with people at home in order to delay cognitive decline and health-related problems, etc.

There are more and more people accepting robots being part of their daily life. According to a survey of accepting social robots in three countries – US (1369 people), German (1382 people) and Japan (1390 people) – performed by Hiroyuki et al [HNI], 60 to 70 percent of the participants report as being "very comfortable" or "somewhat comfortable" when asked if they would accept robots being part of their daily lives. Indeed, a human can develop a social relationship with a robot even when the robot's cognitive, behavioral and interactive capacities are much simpler than those of any human being or any social animals. However, different types of relationships can exist along the spectrum that ranges from treating a robot as a machine to perceiving it as a social and persuasive agent.

1.2 Defining social Robots

Humans have always been developing "social" relationships and get attached to objects in the world around them [Dau03].

Traditionally, robots were machines primarily used in clearly defined environments with specified tasks [Dau03], in particular in manufacturing environments. Humans have been interacting with such robots in the same way they interact with other machines. If any *relationship* exists at all with such robots, then it is the same type of relationship that humans may have with fridges or cars. For example, we often name our car or even talk to it. However, this relationship is not strong and is uni-directional because these objects remain passive, and never initiate interactions. Moving to domestic environments, some kind of service robots such as cleaner robots, another kind of machine-like robots, have capabilities of moving in a home and sweeping up dirt. Although the robots do not have any interactive communicative abilities with humans, they can create *social relations* with some families. For example, some people named the robot, used them in groups of two vacuum robots or sometimes say "excuse me" to the vacuum if he/she bumped into it [FD06] .

These machine-like robots are really different from recent developments of robotics with social skills that can interact with humans through verbal and non-verbal communication in human environments. These *Social Robots* are robots designed to interact with us through social behaviors such as recognizing and engaging humans in conversation or even have strong *affective relations* or *intimate relations* with human and they are expected to help people in many aspects (e.g. physical care, episodic and autobiographic memory). The two next subsections will describe a definition of *Social Robots* and motivation of *Socially Assitive Robot*, a particular type of social robots, aiming at helping people through interactive communication.

1.2.1 Definition & Classification

Bartneck and Forlizzi [BF04] defined a social robot as follows:

"A social robot is an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact"

Like other robots, social robots are physically embodied (avatars or on-screen synthetic social characters are not embodied and thus distinct). However, some robots using a screen to display the robot's head such as JIBO [RMK14] or Buddy [Rob16] can be considered to be social robots because the screen-based head sets expectations of verbal interaction.

Following their definition, a requirement for a social robot is autonomy. Therefore, a remotely controlled robot will not be considered as a social robot because it does not make decisions by itself. However, semi-autonomous robots – with shared control with human pilots – somehow can be defined as a social robot if it interacts with an acceptable set of situated social skills.

Based on the definition, the social robots can be characterized according to several parameters: (1) form (abstract, biomorphic or anthropomorphic), (2) modality (from unimodal to multimodal communication channels), (3) the knowledge about social norms, (4) the degree of autonomy, and (5) interactivity - the potential to exhibit causal behaviors.

Hiroyuki et al [HNI] classified social robots according to function and appearance (see Figure 1.1). For example, from left to right, an Amazon Echo robot has mechanical appearance and performs some practical functions such as checking weather, traffic, user's calendar, etc. Some companion robots such as JIBO, BUDDY are endowed with ability to communicate with human though speech and express some emotions, turning head to people, etc. At the bottom of the figure, robots with human-like appearance such as Pepper and Nao have the ability to express non-verbal behaviors.

1.2.2 Humanoid robots

Comparing with other robots, humanoid robot's behaviors seem to be more attractive because of human-like appearance [Kie+08]. Humanoid robots are believed to be suitable for communicating naturally with human due to their human-like bodies that enable humans to immediately understand their social cues such as gaze, gestures. Therefore, humanoid robots could be used to help people in many communication tasks such as guiding people in public space such as museums, train stations or shops. For example, Hayashi et al [Hay+07] demonstrated a positive scoring of communicative robots at a train station where customers stop to watch humanoid robots greeting and giving some information.

Human even treat humanoid robots as if they treat other people following their culture (e.g. we do not want to interrupt a conversation between two other people). For example,

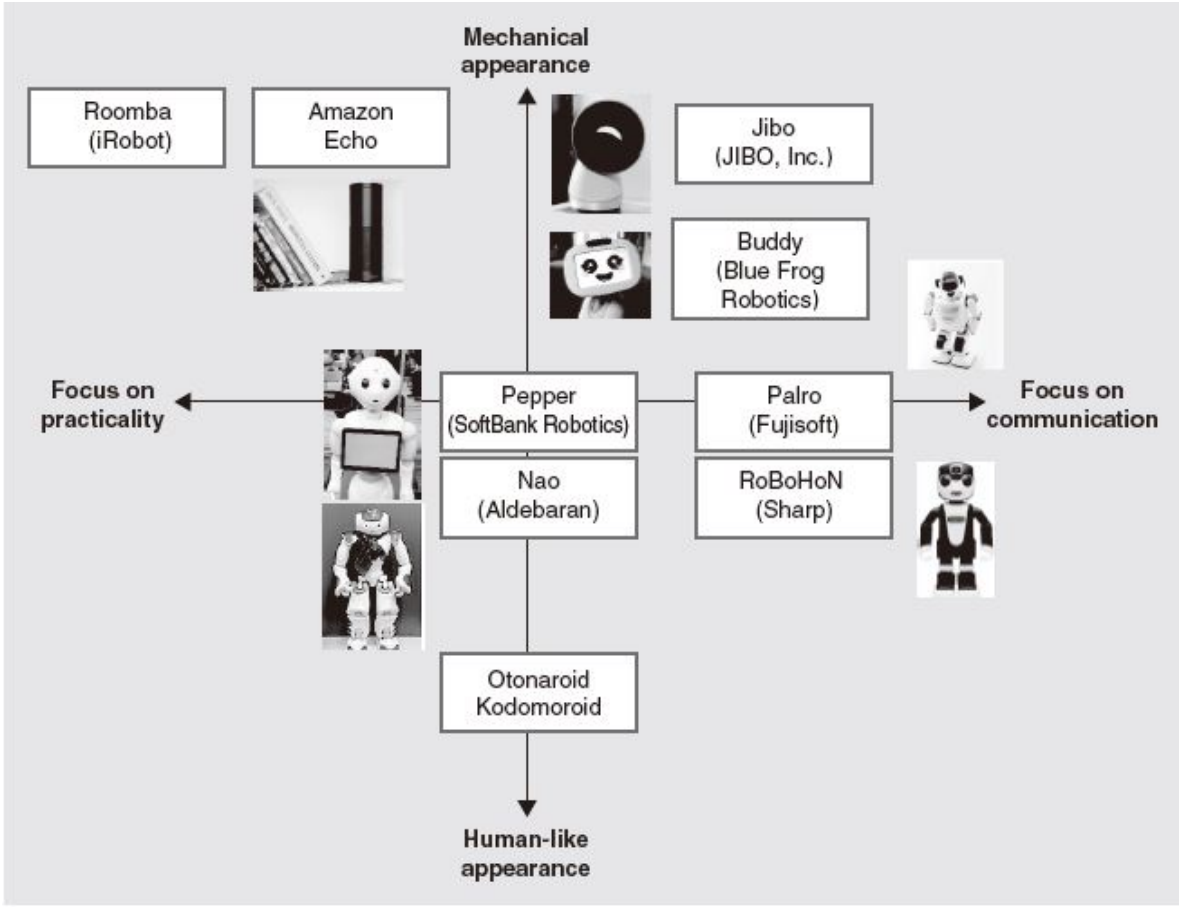


Figure 1.1 – Classifying social robots according to function and appearance [HNI]

Hayashi et al [Hay+07] let two humanoid robots to provide information for passengers in a train station. They found that people were much less likely to interrupt a conversation between the two robots. They suggested that people did not want to be engaged because they felt that the robot couple would be less likely to open the conversation to them. They suggest that people use the same culturally-grounded social norms to decide how to interact with robots.

In a museum, after experiencing a guidance by a humanoid robot, children were interested in more exhibits, especially, items that were introduced by the robot [Shi+06].

Humanoid robots could be more effective than display devices in advertising or providing information. For example, Kanda et al [Kan+09] found that robots provide more useful information and encourage more shopping than display devices in a mall where robot guide and recommend people go to shopping/restaurants. They conclude that the establishment and development of human-robot relationship could increase advertisement effects. Figure 1.2 provides subjective evaluations comparing the efficiency of robots vs. display devices in providing information and interesting in shopping after interacting with them.

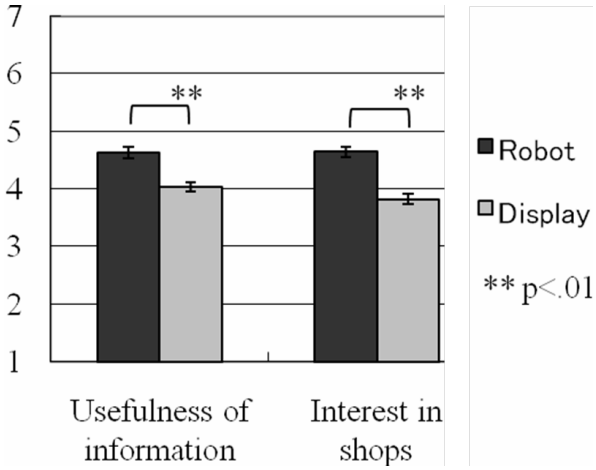


Figure 1.2 – Comparing robot and display in guiding and recommending shopping [Kan+09]

Another advantage of humanoid robots is that we could more easily design and evaluate their social behaviors using established socio-psychological metrics, because they have similar physical morphologies with humans. The problem of mapping perception to behavioral responses could then be simplified [BS02].

We just mention humanoid robots, a specific type of social robots with some useful features. The next sub-section describes several interesting properties of social robots and why they attract both users and researchers.

1.2.3 Why Social Robots as Assistants?

In recent years, there has been an increase in the application of social robots in helping people such as assistant in hospitals, shopping assistant, tele-medicine, hotel service, taking care elderly, helping daily people in daily life at home (household), and educating children. There are many reasons why the robots are expected to be popular to assist people in the applications.

The text below lists some of the reasons:

Engagement Social robots have more engagement abilities than other interactive technologies. Physical robots elicit more favorable social responses than other interactive technologies. In contrast to traditional interactive interfaces such as screen and voice interfaces, social robots provide for an alternative mode of interaction: an embodied interactive experience for users [LHZ17]. Comparing with applications on mobile phones/virtual agents, robotic technologies have stronger social engagements and are more effective reminding users to keep their health-care schedule [RMB14]. In fact, robotic devices could be useful to promote health monitoring and health behavior such as diet and exercise. Especially, a humanoid robot (one kind of social robots) with human-like body movements such as shaking hands, greeting, and pointing could be more likely understood by

adults and children than interaction with an electronic interface such as button or touch screen [Shi+06]. Mann et al [Man+15] showed that embodied robots make people feel more relaxed and respond better comparing with computer tablets used to give health care instructions.

Trust Humans trust a social robot as much as – or even more – they trust a human.

When robots have convincing social kills, people tend to trust them as they trust humans. For example, Kahn et al [KJ+15] suggested that people will form increasingly intimate and trusting psychological relationships with humanoid robots when the robots increase their social interaction abilities. In their study, they compared a 20 minutes interaction of participants in three conditions with: (1) a humanoid robot with a high level of social sophistication (controlled by Wizard of Oz) and (2) a rudimentary social robot and (3) with a human. The three agents performed a lab tour guide. In each condition, the lab tour guide would share a failure (at the end of the interaction) and asked each participant to keep it secret from the experimenter. Their results showed that the majority of the participant kept secrets of the highly social robot (59%) as well as the human (61%) and there is no statistical different between the two conditions. In contrast, the percentage who kept the robot with rudimentary social interaction (11%) is significantly different from the other conditions. In another work, Bethel et al [BSS11] showed that children shared a secret delivered to them by another person, with a humanoid robot as much as they share it with an adult human. The results of these studies suggested that human may trust social robots. Therefore, there is a promise of a "persuasive robot" area, where robots are designed to encourage people to change their behaviors (e.g. robots play roles of weight loss coach [KB07]) or asking for intimate questions about health or sensorimotor skills.

Role Robots can easily exchange their roles with humans. In education applications, robots can play many roles for helping children to learn. In one hand, robots can take many roles such as acting as peer learning companions, tutors or mentors [Mub+13]. For example, robots help students in remembering vocabulary [Sae+10] by playing a game with them and encourage students each time they remember a vocabulary item. In the CoWriter project [Lem+16], a robot plays the role of a *"bad writer"* that is taught by children who were diagnosed with visuo-constructive deficits (e.g. difficulty to write) . By shifting the roles of the child from the *"underperformer"* to *"the one who knows and teach"*, the robot helps children not only practice to write better but also recover their self-confidence.

Companions Animal-like social robots can play roles of pets while avoiding disadvantages of the living animals (e.g. allergies, hygiene, etc). In therapy applications [ASI01], animals can actually help people to significantly reduce their blood pressure, heart rate and anxiety levels, etc.. However, there are some disadvantages of using animals for therapy such as hygiene risks (e.g. infection), fear of animals (e.g. biting), allergies, etc [Sch06]. Animal-like social robots can solve this problem while keeping the advantage of living animals. For example, there are several effective roles [Abd+18] of animal-like robots found in elderly care such as:

Affective therapy Reducing depression, agitation scores and increase quality of life scores of elderly people [Sai+03]; [Wad+03]; [Jør+16]; [VS+15].

Cognitive training Improving aspects of cognition, such as working memory or executive function [Tan+12]; [Kim+13]; [Wad+08]; [Tap09].

Social facilitator Improving sociability between human and robot or between the subjects with other people [KFB09]; [Sab+13]; [KTT06]; [Chu+17]

Companionship Reducing significantly loneliness scores of participants after they interacting with the robots [Kan+03]; [Mac06]; [Rob+13].

Physiological therapy Even the effect of SAR on physiological therapy is less clear, some studies show that social robots can decrease blood pressure when participant interact with the robots longer [RMB15].

1.2.4 Socially Assistive Robot (SAR): Long-Term vs. Short-Term Interactions

SARs can be classified following two operational groups [Abd+18]: (1) service robots and (2) companion robots. Service robots aim at helping people's activities of their daily-life. In contrast, companion robots often play the role of pets or majordomos to improve the psychological status and overall well-being of its users (e.g. Sony AIBO [Kan+03]).

SARs are typically facing two situations with quite different timescales and related challenges:

Long-term interaction . Long-term (LT) interactions often target one single user with the challenge of engaging into open-domain conversations, establishing affecting relation. In other words, the robots service for individual customers and enhance relationships with them.

Short-term interaction . In contrast, short-term (ST) interactive robot could be less required affective relation and more focus on task demands, but they still requires attention.

1.2.4.1 Long-term Interactive Robot

Long-term robots are social robots that used with human in a long period of time (several months, years) or in everyday life. They often share the private like of a few set of users, being able to establish social glue and share common experiences. In order to kept engage with human users, the robots need to be provided with emotional and affective skills.

There are several main applications of LT interactive social robots such as elderly care, autism therapy and education. For examples, one of landmark examples of long-term social robot is Paro, a animal shaped like (with a seal embodiment shown in Figure 1.3 (a)) therapy

(a) Paro, an animal robot used for therapy and household [WS07]

(b) Sony Aibo ERS-210 - a robot dog [Mel+09]

Figure 1.3 – Long-term interaction robot: companion robots

robot, which is used for interacting with patients with three purposes: psychological, physiological and social effects [WS07]. The Paro robot was introduced in a public area at Care House where they interact with elderly people in total 9,5 hours per day. Wada and Shitaba found that people interacting with the robot reduce their stress level and establish strong ties with the robot (e.g. they greet the robot when seeing it again).

One of the most sophisticated long-term interaction robot is Sony's AIBO, also an animal-shaped robotic 'dog' shown in Figure 1.3 (b). The robot was designed to mimic biomechanical motions of dogs with sophisticated adaptation capacities. Melson et al [Mel+09] compared the way children interact with AIBO and found that almost all children (96%) engage into interaction using social movements, e.g. 'offering the ball'. Especially, children offered the ball to the robot dog more often than to the living dog. However, the children were more likely to talk (e.g. give questions) with the live dog than the robot dog. The results showed that children could engage with the robot dog as an believable interactive partner.

Another study on long-term human robot interaction was conducted with a robot Robovie (designed to have safe and stable hardware for interactive communication using gestures, shown in Figure 1.4) in an elderly care center [SKH11]. The robot was placed here for 3.5 months and controlled by a non-robotic staff of the center. The robot gave emotional support to elderly by conversation such as greeting (calling them by name) or encouraging them perform some difficult tasks. The robot also played as role of a child to ask the elderly some question such as "*what is this?*" and the staff confirmed that the elderly perceived the robot as a child. The authors found that the elderly tend to interact with robot, even though at the first time they were not sure how to approach the robot.

For autism therapy, Bharatharaj et al used KiliRo [Bha+17], a Parrot robot (shown in Figure 1.5), to improve the social interaction abilities of children with autism spectrum disorders

(a) Robovie used in an elderly care center to greet people by name and played as a child to communicate with the elderly [SKH11]

(b) Robovie interacting with children in elementary school [Kan+04]

Figure 1.4 – Robovie Robots

without risks of being harmed by real parrots such as biting.

For education, Kanda et al [Kan+04] used two Robovies (provided with many interactive behaviosr such as hugging, shaking hands, exercising, greeting, etc.; each arm of robot have four degrees of freedom requiring minimum torque motor to control so that it could be used easily and physically stop in case of dangerous situations [KI16]). The robots interacted with children at elementary school to teach them English language during two weeks. The research found that most of children did stop interacting with the robot after the first week because they had high expectation from robot behaviors and were disappointed. However, the rest of children who kept interacting with the robot got higher English scores. Because the improved scores just show after the second week, they suggest that the robot's influence will depend on its ability to create a relationship with the user. The children interaction also gradually reduced, especially during the second week. That means the robot fail to maintain long-term relations with humans. They suggested this was because of the body and appearance of the robot. They also recommended that the duration of interaction could be increased if the robot possesses a humanoid body (the iCub-humanoid used in my work could be expected to make the robot more interesting due to its human like appearance with mouth articulation as shown in Figure 1.8)

1.2.4.2 Short-term Interactive Robots

Short-term interactions are typically task-oriented, repetitive and usually performed with many users in a short period of time. The robots often perform tasks in a professional environement such as welcoming a client, giving directions, conducting interviews etc. The short-term robot should cope with a large variety of user profiles and be able to adapt in a very short period of time.

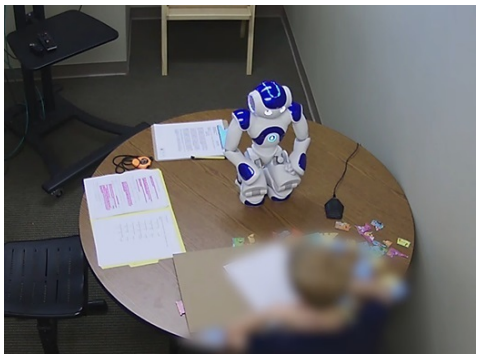
Figure 1.5 – KiliRo, a parot robot, interacting with autism children [Bha+17].

For example, Foster et al [FKL14] built a bartender robot, named JAMES, interacting with multiple customers for serving drinks to them in a bar. They actually learn action selection policy for the robot from simulated environment with Markov Decision Prossess (MDP) models that map state features to actions so that to maximize expected cumulative reward. The robot was provided with a limited set of actions including dialogue acts for clarifying drink order (e.g. "Did you say 'blue lemonade'?").

Bethe et al [Bet+16] used Nao robot to interview children whether they were bullying victim or not. Their hypothesis is that children would be more likely reporting bullying to a robot than human interviewer. However, they found that the are no significant differences between human and robot interviewers rated by their parents in terms of niceness, trust, helpfulness, discomfort, etc.

The Robovie robot was also used to interact with visitors and provide them with explanation about items in a museum [Shi+06]. The visitors were provided a radio frequency identification (RFID) tag, a technology which enables the robot to easily identify individuals and get visitors' personal information, times of registration/return tags, when the visitor approach the particular exhibits and so on. Therefore, the robot can greet visitors by name or wish them a happy birth day, say goodbye to the departing visitors, etc. With the ability of calling users by their names, the robot make the users feel more friendly and can be more affected. By questioning participants, they found that most of people feel robot interesting and friendly and just few of people reported anxiety about robot's interactions and future robots. Based on feedback from visitors, they found that after guided by robot, children seems developed an interest in new exhibits. However, there were few children being afraid or not caring about interacting with robots.

(a) Robot bartender, namely JAMES [FKL14].



(b) Nao, a humanoid robot interviewing bullying with a child.

Figure 1.6 – Examples of short-term robot interaction

Figure 1.7 – The robot is chatting with custom, recommending them shopings/ restaurants and using deictic gestures to guiding them in a mall [Kan+09].

Humanoid robots could also be used to guide customers in a mall and recommend them go to shops/restaurants [Kan+09]. When the robot detects a person near the robot, it greets the person. The robot can chat with the person and ask his/her preference and offer route guidance or shopping information of the day. The robot used some nonverbal behaviors such as deistic gestures for giving direction to the custom as shown in Figure 1.7. They found that robot could encourage shopping and is rated more useful than display devices in the mall.

1.2.5 Our work

Even though short-term robots require less affective relation and are more focusing on task-oriented behaviors, they still require to monitor attention and engagement. Our work focuses on the development of socio-communicative abilities for enabling a humanoid robot to perform short-term interactions. Particularly, we build here a learning framework for providing a

humanoid robot with multimodal interactive behaviors – using speech, gaze, arm gestures, etc.

The target scenario is a neuropsychological interview with an elderly person. A corpus recorded with a professional interviewer will serve as demonstration data.

1.3 Learning Methods

One of the most advantage of "semi-autonomous" robots is that they have the ability to actively "learn" about themselves and their surrounding world, which heavily distinguishes them from traditional computing technology [AS05].

One way in the direction of embodied cognition and socialization is to let robots train via developmental learning: in a process similar to humans, the robot will improve its control system through interaction over time, acquiring new skills through interaction with its environment. This process is often ruled by a try-and-error policy, necessitating clear rewarding procedures and pre-existing operation modules for decoding intentions of others.

Alternatively, SOMBRERO (described below) builds on immersive teleoperation and explores the possibility for robots to learn optimal cognitive and social behaviors via demonstrations performed by humans.

1.3.1 Developmental Robotics

In the "developmental" approach, the robot learns by accumulating knowledge and skills through self-experience. The aim of developmental robots is to allow robots to learn new skills and new knowledge in life-long and open-ended interaction via autonomous exploration of the world and social interactions with caregivers.

Mohammad et al [MN15] proposed an approach to endow social robots with interactive skills through three stages:

interaction babbling the robot learns basic sensorimotor skills.

interaction structure learning the robot uses the learned basic skills to learn a hierarchy of probabilistic/ dynamical systems.

interactive adaptation the robot engages in HRI to adapt the hierarchical model to different social situations and partners.

The term *developmental* refers here to these stages where the robot accumulates basic skills by itself. Then, the basic skills are used to achieve the higher stages. This requires actual engagement or motivation – also termed as "curiosity" in the literature – in interactions to achieve any progress in learning.

The key idea of the learning developmental robot lies on intrinsic motivation (IM) that enables the robot for self-directed exploration, automatically learn features of environment for its inherent satisfaction [OK09]. For example, Castro et al [CG+14] defined several motivations for for a social robot (named Maggie) (e.g. **energy** is motivation of **survival**, **loneliness** is motivation of **social**, etc.). Then the robot learned to decide which action is to deal with each motivation (e.g. if the robot perceives low energy, the motivation of survival is augmented so that the robot go to charge at a docking station; if there is a person are close to the robot, social motivation is strengthen, then the robot tends to interact with the person). As an other example, Breazeal et al [Bre04] built a robot head (named Kismet) aiming at developing social intelligence by interacting with people as if they teach an infant. By providing Kismet with a *Theory of Mind* and a synthetic nervous system, it pro-actively engages in social exchanges with the caregiver and to acquire interesting and relevant information from the environment.

The concept of the developmental robotics is really difficult to apply to professional skills, especially for running multimodal interactions with humans since it requires a lot of motivations and social rewards that are difficult to implement. Exposing people to true negatives – i.e. failures of its social behaviors – is also illicit. Developmental robotics requires the accumulation of many basic skills like a child before it can learn high-level skills.

Therefore, in order for a robot to perform professional interactive tasks (e.g. playing roles of an neuro-psychologist interacting with Alzheimer patient described in the next chapter), we study another approach – termed as *Learning by Demonstration* (described bellow) – to endow our robot with multimodal socio-communicative behaviors.

1.3.2 Learning by Demonstration

Learning from demonstration (LfD) aims at enabling non-robotic experts to teach a robot new skills without requiring professional robotic backgrounds. The demonstration can ensure that the robot will directly learn and control actions via its sensori-motor abilities. Demonstrations are often used to limits searching space of Reinforcement Learning [BG13].

There are various ways to train robots from demonstrations that could be classified in low vs. high level.

1.3.2.1 Demonstrating Low-level skills

Low-level learning refers to the terms of *skill*, *motor skill*, or *primitive action*. The goal of this kind of learning is to build a library of primitive actions that could be used in some specific tasks. For example Pastor et al [Pas+11] learn a PR2 robot to play with a pool stroke and manipulate a box with chopsticks.

Learning robot motor skills could be done using **supervised methods** such as Neural Networks, Hidden Markov Model or Gaussian Mixture Regression. For example, Asada et al [Asa90] used a feed-forward neural network to map a measured force with a corrected tra-

jectory. Sofiane et al [Sof+14] also used a feed-forward neural network to enable a humanoid robot learns different postures by associating what it did with what it saw (the robot first produced a random predefined posture, then see how a teacher mimics the posture). Rahmatizadeh et al [Rah+16] used Long-Short Term Memory (LSTM) layers combined with Mixture Density Networks to train robot arm to pick and place boxes in a virtual environment. Furthermore, there are several dynamic-based model methods are used. For example, Pastor et al [Pas+11] proposed to use the *Dynamic Movement Primitive (DMP)* framework to generate varieties of elementary movements while controlling via -points. The key idea of the DMP method is based on the assumption that complex movements may be decomposed into a set of primitive movements executed in sequence and/or in parallel. Motor skills also could be learned by reinforcement learning (RL) combined with DMP. For example, Kormushev et al [KCC10] developed a compact framework with two consecutive steps: (1) using DMP to learn primitive actions, then, (2) using Reinforcement Learning to refine the coordination between the set of primitives. They succeed in performing *Reaching* and *Pancake-Flipping* tasks with an 7-DOF arm robot [PS08].

1.3.2.2 Learning high-level behaviors - multimodal interactive behaviors

While learning low-level tasks focus on trajectories of primitive actions, the learning high-level tasks concentrates on how to textbfalign/coordinate these actions to fulfill an elaborated task. Especially, in the context of social robots, learning high-level behaviors can be consider as learning **multimodal interactive behaviors** – nonverbal behaviors such as arm gestures, head movement, etc. that coordinate with speech activities – so that to conduct a successful dialog.

Teaching robots such interactive behaviors is more complex than teaching them the low-level skills (object related skills) because of the inherent ambiguity of the nonverbal behaviors, which depends on many factors such as the social context, cultural and personal traits, etc. [MN15]. The social behaviors of robots here do not only refer to endogenous actions/movements for achieving tasks but also to reactive behaviors when interacting with humans.

1.3.2.3 Incremental vs. Batch Learning

Incremental learning is a process that enables a robot to learn when performing the task.

Most of on-line training methods teach robots with low-level skills. For example, Ribsky et al [Ryb+07] alternates between dialogs and demonstrations for teaching robot to perform some simple tasks like picking up a block. They train the robot in 2 phases: (1) LearnTask and (2) FollowLearnTask. The first phase checks if the state of the environment is satisfactory. Then, the second phase triggers the robot following human command. All of tasks and learning methods are hand-coded by rules and encoded by a directed graphical model. Similarly, Qureshi et al [Qur+16] used deep Q-learning to enable robot to learn social interaction skills

by interacting directly with people rather than by imitating human partners.

Active learning is a specific type of incremental learning in which robots collect **labels** during the learning process by generating actions that impose the learning performance. In particular, the robot with active learning becomes more active to provide feedback to human teacher. As an example, this method enables the robot to ask questions to the teacher in order to obtain labels for unlabeled training data. As a result, the teacher could save time by giving to the robot the most efficient information.

In contrast, batch learning consists in providing all of training data at once and producing models.

1.4 The SOMBRERO Framework

Our works focus on the development of sociocommunicative abilities of a SAR for **short-term interactions**. In this section, we present the SOMBRERO framework which aims at providing a humanoid robot with multimodal interactive behaviors – such as speech, gaze arm gestures, etc. – in order to perform a neuropsychological test, demonstrated by professionals.

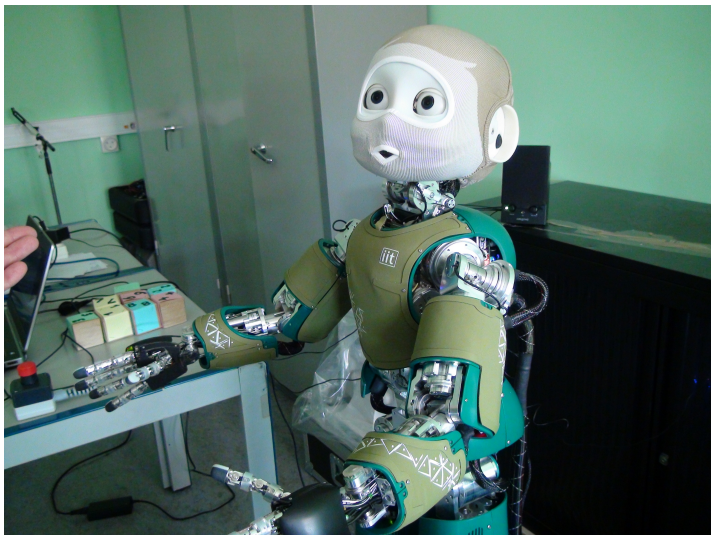


Figure 1.8 – *Nina*, the iCub2 humanoid robot with mouth and lips articulation which has been used in this work.

In the SOMBRERO approach, learning framework is performed in three main steps illustrated in Figure 1.9. Firstly, we collect representative interactive behaviors from human tutors especially by professional coaches. Secondly, comprehensive models of interactive human behaviors are trained from the collected data with considering a priori knowledge of users' models and task decomposition. Finally, *gesture controllers* are build in order to execute the desired behaviors by the target robot.

We proposes two strategies to collect interactive data:

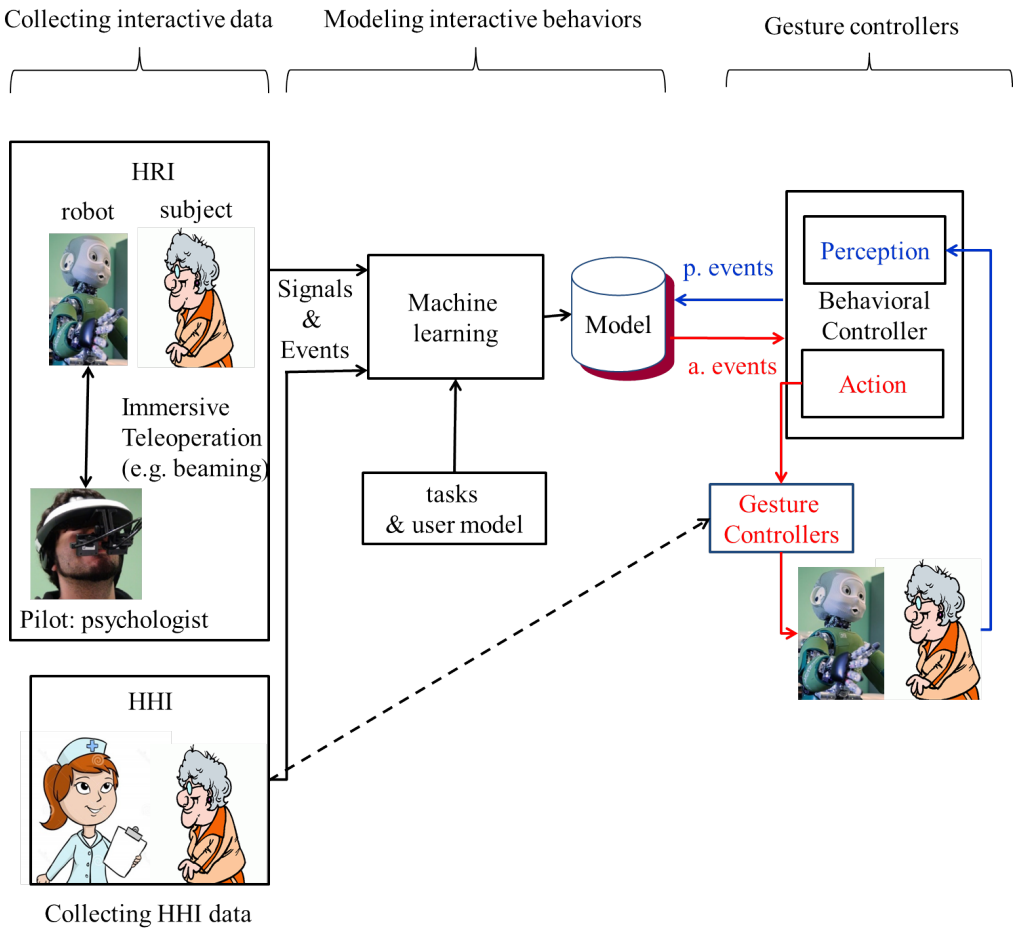


Figure 1.9 – The three main steps of learning interaction by demonstration: collecting interactive data, learning a behavioral model and building appropriate sensorimotor controllers. The collection data could be drawn from HHI or HRI data. In our work, it is necessary to collect HHI for designing HRI with immersive teleoperation as well as gesture controllers. HHI is also used for designing baseline multimodal interactive behavioral models.

HHI Firstly, HHI data are collected to design robot actions. Based on the data, we analyze and model primitive actions that robot is able to mimic. Here, the scenario can be redesigned so that to overcome limitations of our humanoid robot (such as taking notes on a paper notebook). At that time, gesture controllers are also built according to the action requirements for the adapted scenario.

This first strategy is necessary for designing gesture controllers and the data could be used to train multimodal interactive behavior models for the humanoid robot. However, HHI data are not easily applicable to the robot because of its limited both perceptuo-motor abilities compared with humans.

Beaming The second strategy consists in endowing humanoid robots with cognitive, emotional and social skills via immersive teleoperation by human pilots. This technique

called *Beaming* [Gui+15] allowing the human pilot to perceive, analyze and interact with a remote environment through a robot embodiment. The work is based on a hypothesis that the human operator will deal with both scaling and social problems by optimally exploiting the robot's affordances. This is due to the fact that humans are much better in performing social interactions than in engineering social models. Instead of transferring Human-Human Interaction (HHI) to HRI by adapting these behaviors to the robot, this strategy solve this problem by 2 steps: first using beaming technique to deal with technological, psychological and sociological constraints and then using machine learning to model interactive behavioral model directly on the HRI data. Therefore, by shaping pilots' perception and action skills through a robotic embodiment, the framework provides a simple way to study the social acceptance and usage profiles of robots without autonomous reasoning, scene understanding and action planning. The *scaling actions from human-human interaction to human-robot interaction*, therefore, is implicitly solved.

This strategy, at least, guarantees that the robot actions could be performed. However, it does not as certain that the perception modules can significantly be transferred because the perceptual limitation of the robot are not feedback to the pilot during HRI.

In order to solve this later problem, semi-autonomous shared control [YD02]; [CG02] between the pilot and autonomous robot could be used. By enabling the pilot to control part-by-part the robot while giving him/her the perceptional feedback that actually follows the perception states of the robot, we can transfer smoothly both perception and action from beaming control to the autonomous robot. With such semi-autonomous robots, we can enable robot to learn incrementally behaviors.

1.5 Summary

In this chapter, we present motivation and application of SAR and the necessity to endow humanoid robots with multimodal interactive behaviors. We present some approaches to learn social behaviors for robots and why we choose Learning from Demonstration for some situated professional interactions. We also briefly introduce our SOMBRERO framework to collect human-robot interaction data that could be used to train multimodal interactive models.

In the following chapter, we will present how we designed interaction scenarios studied in our work and collect/process human-human interactive (HHI) data of these specific scenarios.

Chapter 2

Human-Human Interactive Data: experimental design, acquisition, annotation & characterization

Face-to-face communication is one of the most natural and effective form of human communication. We use multiple modalities such as speech, body, head, arm movements, gaze and make facial expression in communication with each other in daily life. As stated in the previous chapter, we study in this thesis short-term task-oriented face-to-face interactions. Furthermore, the tasks we have studied are designed so that to focus on low-level cognitive resources such as mutual attention rather than high-level cognitive functions that involve reasoning or perspective-taking: the tasks are easily described as finite-state automata, the roles and objectives of each agent are clearly defined at the start of the interactions, etc. Tasks are also quite repetitive, enabling us to implicitly control the statistical coverage of the free parameters of the tasks such as locations of objects, task ordering, etc.

In our work, we processed data collected during two interaction scenarios, especially designed to confront our modeling frameworks with particular challenges:

Put That There (PTT) is a collaborative game focusing on multimodal deixis. An elementary game last around 2/3 minutes: an instructor and a manipulator have to collaborate in order to reproduce a given layout of cubes. The layout is only known to the instructor while the manipulator can move and position the cubes. The task thus requires the coordination of verbal and nonverbal deictic behaviors – such as head and gaze movements, arm pointing, speech – in order to reach effective and efficient collaboration with the manipulator. If the instructions are clear and timely delivered, the manipulator's task is rather straightforward and instructor's perception is mainly dedicated to the monitoring of his/her moves for triggering the next instruction.

Free and Cued Reminding test (RL/RI) is a more complex collaborative interaction focusing on mutual attention and encouragement. This interaction scenario resumes a neuro-psychological test used to assess episodic memory: the RL/RI test with 16 items [Lin+04] is an adpated version of the Free and Cued Selective Reminding Test [Bus84] that professional psychologists use to evaluate the memory and diagnose some potential loss of the episodic memory of patients. The interaction scenario requires the instructor

– the psychologist – to be endowed with particular verbal and non-verbal behaviors for maintaining engagement and encouraging the patients in word retrieval.

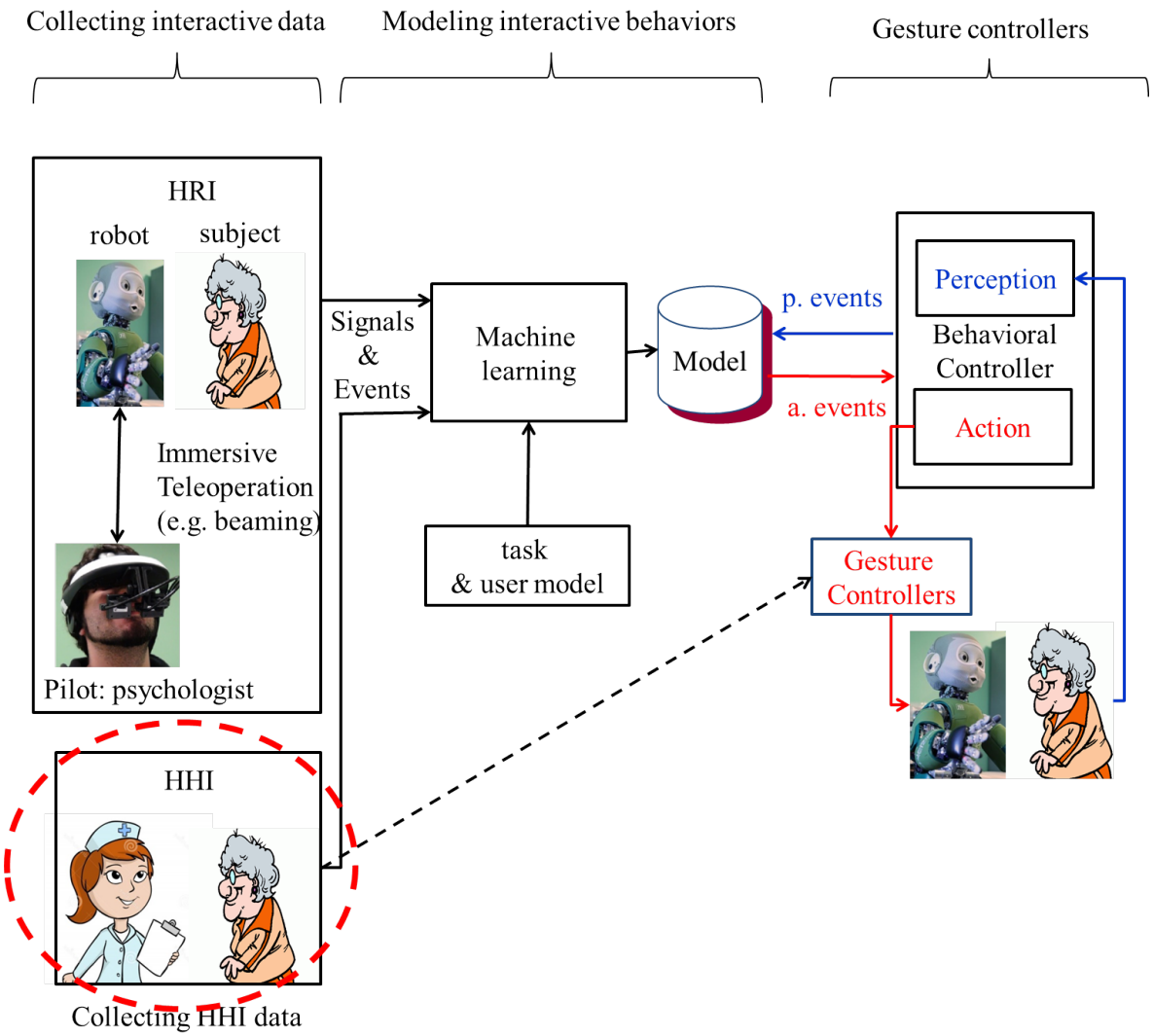


Figure 2.1 – Collect representative interactive behaviors from human coaches in HHI scenario

This chapter presents how these experiments have been designed to collect relevant HHI signals (e.g. speech, gaze, arm of head motion, etc.) in both scenarios. The collection data process is the initial (and crucial) step of the learning framework as shown in Figure 2.1. The data should provide useful features for further training of the multimodal interactive behavioral models (see chapter 3). Here, we are interesting in coordinating both mid-level behaviors (abstract behaviors such as naming, arm pointing, fixating particular objects) and low-level skills (raw signals such as trajectory of head movements). For mid-level behavioral features, raw signals are processed (often via semi-automatic annotation) to get discrete events, organized in multimodal scores. These HHI scores can be easily played back by our humanoid

robot with few adaptations to emulate HRI (see chapter 3).

2.1 "Put That There" (PTT) data

2.1.1 Scenario

The PTT dataset[1] has been collected by Mihoub et al [Mih+16]. This face-to-face interaction involves an instructor and a manipulator who performed a collaborative task called "Put That There". The experimental setting is shown in Figure 2.3. In this scenario, the manipulator should move cubes from a reservoir close to him to a central chessboard, following instructions given by the instructor. The instructor is the only one to know the target arrangement of the cubes, while the manipulator is the only one being able to move the cubes (see Figure 2.2).

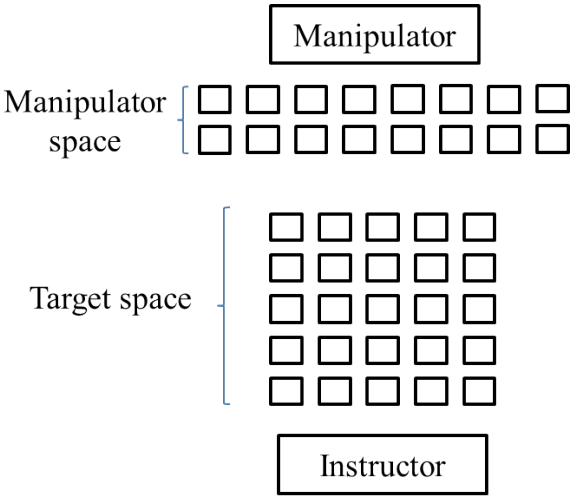


Figure 2.2 – Table of "Put That There" scenario.

The initial arrangement of cubes in the reservoir and the set of moves for each game are pre-computed so that deitic gestures are equally distributed among cubes and locations. These pre-computed instructions are therefore given to the instructor on demand via a tablet placed in front of the instructor and which screen is only visible to him/her: the instructor can browse a *.pdf* file that displays the elementary updates of the chessboard.

Therefore, this task requires the instructor and manipulator to cooperate: share knowledge and coordinate their sensory-motor abilities. Each of the instructor/manipulator dyads performed 10 games consisting in reproducing a target arrangement of ten out of sixteen cubes, with an implicit control of the gaze and hand gestures thanks to the initial and final disposition of the cubes. This balanced statistical coverage of behaviors – via the pre-computing instructions – provides an interesting benchmark to collect human strategies used to maintain

[1] `http://www.gipsa-lab.fr/projet/SOMBRERO/data.html`

mutual attention and coordinate multimodal deixis (finger pointing, head, gaze, etc.) towards objects and locations.

Figure 2.3 – First-person view of the interaction captured from the instructor's head-mounted scene camera. At a game onset, the cube reservoir close to the manipulator is full (16 cubes). The instructor then asks the manipulator to put given cubes at certain places on the central chessboard: at the center at game onset then left/right/on top/at the bottom of cubes already moved. On the figure, the current eye fixation (point of interest) is depicted by a circle [Mih+16]. Note that the tablet on the left hand side of the instructor only timestamps videos

2.1.2 Data Acquisition

HHI data will be used to analyze and train multimodal interactive behavioral models that can generate actions of the instructor as well as to design primitive actions for the humanoid robot. The models are then used to generate actions for the iCub humanoid robot which plays the role of the instructor. Therefore, the collected data should include the multimodal score of the instructor's actions, but the behaviors of the manipulator – the instructor's percepts – should be extracted from the instructor's viewpoint.

The interactive data are collected by:

Motion capture A Qualysis® Motion Capture system (MoCap) monitors motions of instructor's head and right arm (as illustrated in Figure 2.4). The instructor wore a helmet, to which 5 reflective markers are glued in order to capture head movements. His/her arm gestures were detected by 5 other reflective markers glued on his/here right hand and index fingers. The MoCap system consists of 4 infrared cameras facing the instructor.

Eye camera A head-mounted monocular Pertech® eyetracker includes: (1) a eye camera providing gaze fixation data at 25Hz; and (2) a scene camera providing the corresponding

point of interest in the scene for visual perception and annotation. This first-person view provided by the scene camera is shown in Figure 2.3.

Microphone A head-mounted microphone used to record the instructor's speech.

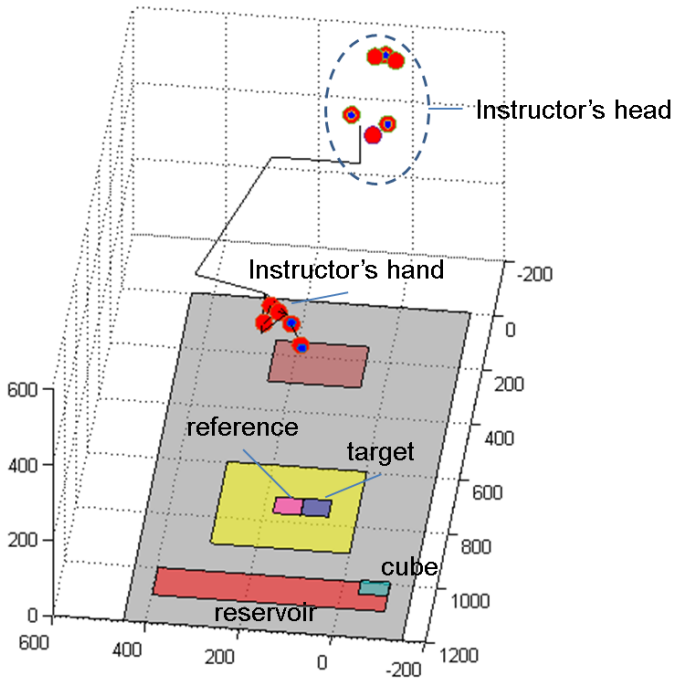


Figure 2.4 – The instructor's head and right arm movements are monitored by the MoCap system: red points cue positions of the reflective markers. The instructor browses instructions on the tablet figured by a mauve rectangle placed in front of him.

The observations also include the three continuous motions of the instructor's head (converted to Euler angles: pitch (H1), roll (H2), and yaw (H3)) automatically delivered by the mocap system.

Those data will be used as ground truth for modeling interactive models that generate gaze, head motions and hand movements of the instructor given his speech and hand movements of the manipulator.

2.1.3 Data Annotation

The data here include 30 games performed by one instructor and played with 3 different partners. For each game, the dyad has to replicate an arrangement of 10 cubes on the chessboard

from an initial random layout in the reservoir. The mean duration of a game is around 80 seconds. The total duration of recorded interactive data is about 30 minutes.

All raw streams are re-sampled at 25Hz. Additional annotations were performed using Elan [HU04] and Praat [Boe+02]. The final observations consist of 5 streams of discrete variables:

Instructor's speech (SP) Audio signals collected by the head mounted microphone are first aligned with text. This first alignement is hand-corrected (hesitations, false-starts if any are then added). Speech of the instructor is further segmented into 6 parts-of-speech: *manipulated cube*, *reference cube*, *relative positioning*, *else*, *none*.

Instructor's arm gesture (GT) The instructor right arm movements are segmented semi-automatically by detecting strokes and labelled in accordance with speech annotation as shown in Figure 2.5. In fact, the maximum velocity at each movement onset is often aligned with the corresponding onset of part-of-speech. There are 5 regions of interest pointed by the instructor's index: *manipulated cube*, *target location*, *reference cube*, *rest* or *none*

Instructor's gaze (FX) The region of interest fixated by the instructor's gaze provided by the head mounted eye camera are labelled with 7 values: *manipulator's face*, *reservoir*, *task space*, *manipulated cube*, *target location*, *reference cube*, *tablet*, *else*.

Manipulator's arm gestures(MP) Manipulator's arm gestures are manually annotation from the scene videos with 5 values: *rest*, *grasp*, *manipulate*, *put*, *none*

Interaction Units (IU) Each game is further segmented into interaction units – that could be also termed as elementary skills or sub-tasks – describing the sequential organization of a repetitive elementary interaction. Interaction Units are distinguished between 6 different values mirroring the activities of the instructor: *get instruction from tablet*, *seek the cube to be manipulated*, *point the cube*, *indicate target position of the cube*, *check the manipulation* and *validate the result*. These IU pace the activities of both agents that are characterized by the about observations.

The challenge is to predict the instructor's co-verbal gestures GT and FX given his verbal activity (SP) and the interlocutor's gestures (MP). The behavioral models (proposed in chapter 3) should thus generate endogenous co-verbal behaviors from endogenous verbal behaviors and exogenous percepts. These models should capture the co-correlation between all of the modalities some of them are both intra-coordination (the relation of modalities that insides the instructor: for example, his gaze (FX) looking the cube, then his arm's pointing gestures (GT) to the cube, and his speech indicates the cube) and inter-coordination (the relation between the instructor's modalities and manipulator's modalities: for example, the gaze of instructors should be driven by arm movements of manipulator to verify the position of target cube). The interactive data should provide rich information of the micro-coordinations (in both intra vs. inter) among the modalities so that the interactive models can captures the coordinations.

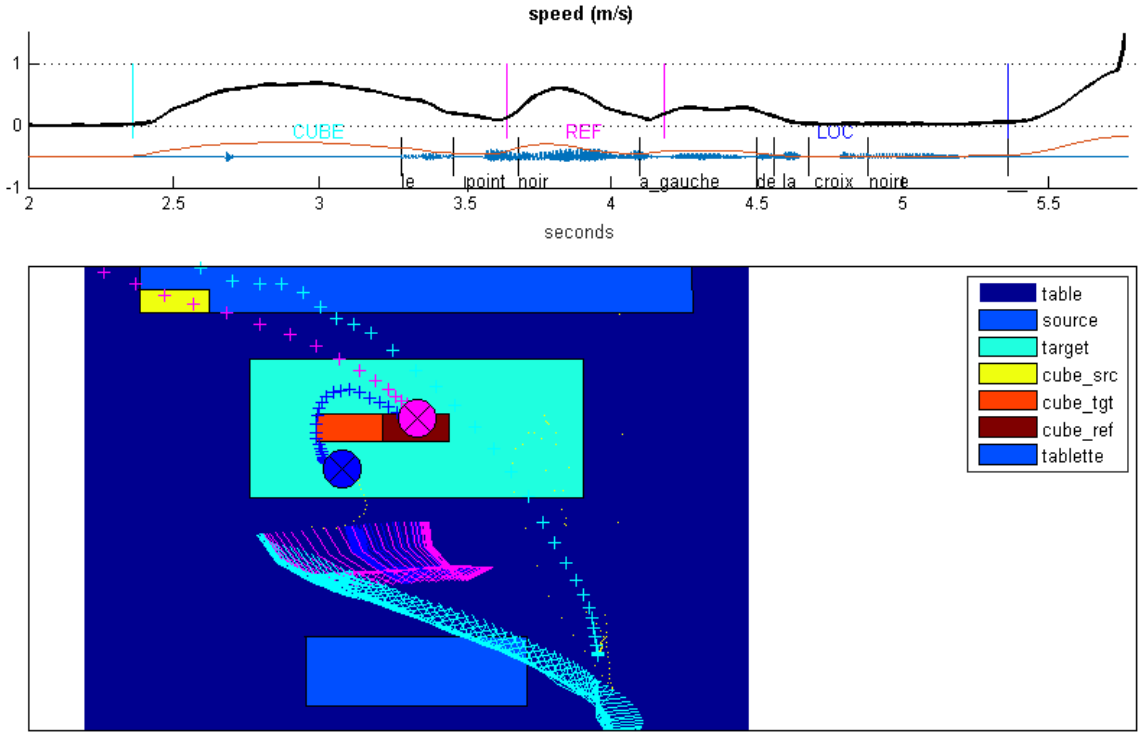


Figure 2.5 – Semi-automatic segmentation of arm movements according to speech and target of the pointing gesture (computed as the intersection (crosses) of the direction of the extended index (lines connecting the 5 landmarks of the hand) with the workbench. The instructor sits at the bottom.

2.1.4 Comments

If onsets of hand gesture are often synchronized with speech onsets of referred locations, we found that speech offsets are “waiting” for the gesture to be performed (see figure 2.5). In many cases, the speech is pausing in order for the gaze to find the location the instructor has in mind and to accomplish the deictic gesture. Speech production – waiting for current co-verbal action to end or next co-verbal action to be planned – is delayed via different strategies: final syllabic lengthening, pausing as well as the production of hesitation ("euh"). These startegies are co-occuring in case of large waiting time (see figure 2.6).

This means the speech, arm movements and gaze are interdependent with each other: while speech often orchestrate co-verbal behaviors, these behaviors – that can be slower than speech articulation because of inertia or cognitive constraints – can in turn influence speech production. This complex relationships of the modalities is a big challenge of building *incremental* multimodal behavioral models.

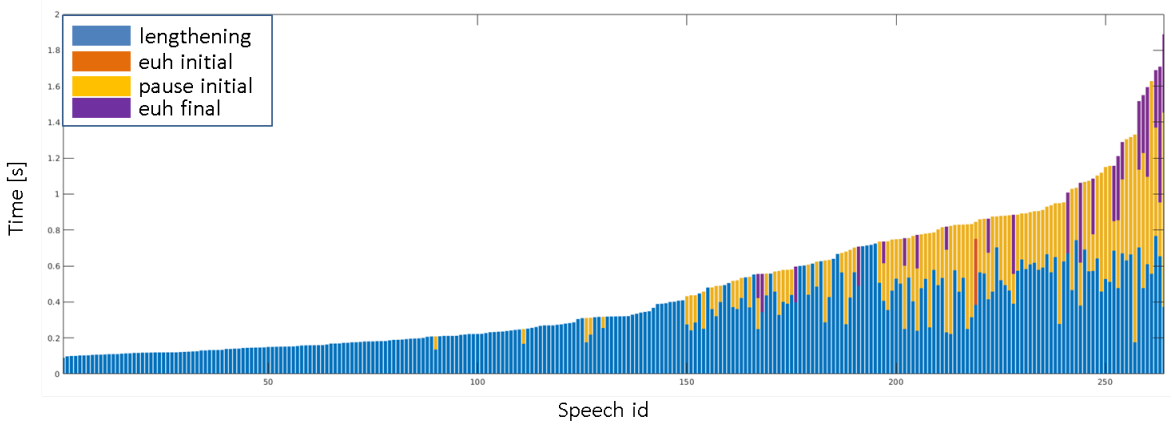


Figure 2.6 – Speech production waiting for arm movement planning: the last syllable of instructor's parts-of-speech cueing cubes or locations is often lengthened in order to wait for current co-verbal action to end or next co-verbal action to be planned. This waiting time is performed via different strategies: final syllabic lengthening (blue), pausing (yellow) as well as the production of hesitation ("euh" in cyan).

2.2 Selective Reminding Test data (RL/RI)

2.2.1 Scenario

The second task that our robot will be involved in is acting like a neuro-psychologist that interacts with elderly people for performing a memory test. This short-term interactive scenario is a French adaptation of the Selective Reminding Test, so-called RL/RI 16 (rappel libre/rappel indicé, avec 16 mots) [Dio+15]. It is often used to diagnose early loss of episodic memory. The test includes four phases:

Learning The subject is instructed to memorize a set of 16 words (e.g. apple) at the same time as their semantic categories (e.g. fruit) (aka learning). The words are learned 4 by 4.

Testing The interviewer tests the recall capability of the subject by asking him/her to spell out as many words as possible either freely or with the hint of their categories.

Recognition Subjects are asked to recognize the learned words amongst a list with outliers

Distractive task Cognitive load is increased by asking the subject to periodically perform a distractive task (such as reverse counting).

In the *Learning* phase, the subject learns 16 words, four by four. This phase consists of 2 tasks:

Identification (the interviewer shows each time 4 items and asks the subject to speak

out loud the items corresponding to the requested categories)

Immediate recall (after the 4 items are identified, the interviewer hides all items and asks the subject to spell out those items again; if a item is missing, the interview will hint by giving its category).

The *Test* phase includes three successive recall tasks separated by a distractive task (reverse counting). Each test consists of a free recall (the subject freely recalls as many items as possible) followed by an indexed-by-category recall (missing items will be recalled by cueing their distinctive categories, such as furniture, fruit etc).

The final *Recognition* phase is a recognition task in which the subject should identify the 16 learned items spoiled by 32 distractors (16 words with the same semantic categories and 16 words of different semantic categories). The interviewer reports answers on a score sheet.

Figure 2.7 – Capturing the multimodal behavior of the human tutor during HHI. Movements of the upper limbs (head, arms and hands) are monitored by tracking 22 markers glued on these segments with a Qualysis ® mocap system. Gaze was tracked using a Pertech ® head-mounted eye tracker.

2.2.2 Data Acquisition

Interactive data of this scenario were collected by Bailly et al [Bai+16]. The interviews were conducted by one unique interviewer – so that subject-adaptive behavior remains consistent across multiple interactions with 5 different subjects. Because the interactive data will serve as demonstration for our iCub humanoid robot, therefore, most of the signals are captured from the interviewer's perspective and there are thus no invasive sensors placed on the subjects.

The experiment setup was almost similar with collecting interactive data in the *Put That There* scenario. We are interested in the multiple modalities of interactions including head motion, arm movements, gaze, and speech. The motion of 25 reflective markers glued on the

plexus, shoulders, head, arms, indexes and thumbs of the professional interviewer was monitored thanks to a Qualysis® system with 4 cameras. The interviewer also wore a Pertech® head-mounted monocular eye tracker that monitors her gaze (see Figure 2.7). Speech data were captured via OKMII high-quality ear microphones and are recorded synchronously with a side-view video by an HD camera (see Figure 2.7).

2.2.3 Annotation

HHI demonstrations were performed by a female professional psychologist (the robot's teacher), whose behaviors our robot will imitate. We collected her multimodal behavior (speech, head movement, arm gestures and gaze, see figure 2.7) when interviewing five different elderly patients as well as the speech of the interviewees. These continuous signals were then semi-automatically converted into time-stamped events using Elan [Wit+06] and Praat [Boe+02] editors.

With Elan (shown in Figure 2.8), we basically determined hand strokes triggered by the interviewer to grasp and act on resources (workbook, notebook, chronometer) and regions of interest for fixations. With Praat, we hand-checked the phonetic alignment performed by an automatic speech recognition system and added prosodic annotations as well as special phonetic events related to backchannels and breath noises. Some values of each labeled modalities are shown in Table 2.1.

Table 2.1 – Semi-active labeling discrete events

Modalities	*Values*
Gaze target	subject face, book items (subject tablet) and score sheet (interview tablet)
Arm gesture	preparing scoring, scoring, showing/hiding items, rest
Speech	turns, words, sentences
Task execution	sub-tasks such as item identification, immediate recall, counting, free recall,etc.
Backchannel	yes, no

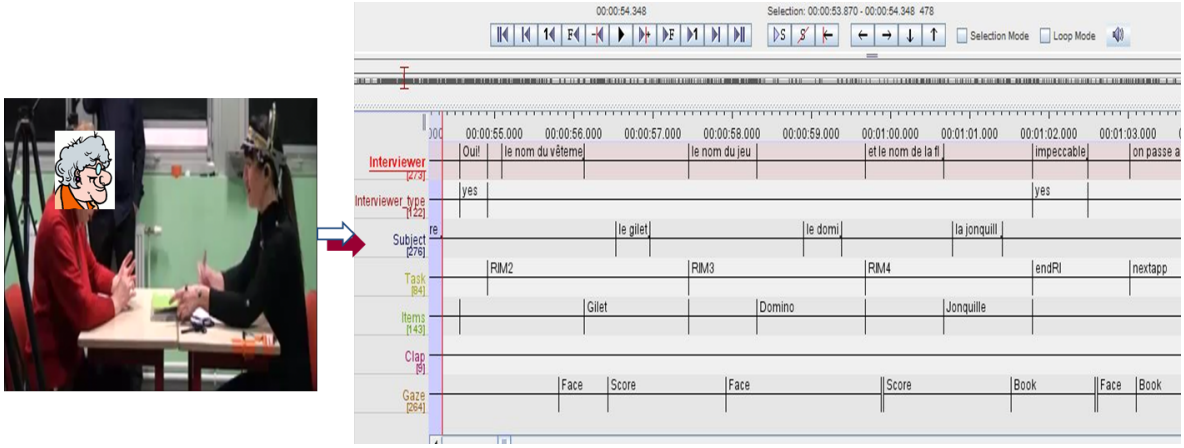


Figure 2.8 – Semi-automatically annotated data with Elan software.

This HHI multimodal data thus consists in time-stamped phones, head/arm/hand gestures and gaze events. We then developed modality-specific gesture controllers to map these events to robotic actions that a human observer could perceive with the correct semantics. HHI to HRI re-targeting is thus performed using multimodal events as pivots.

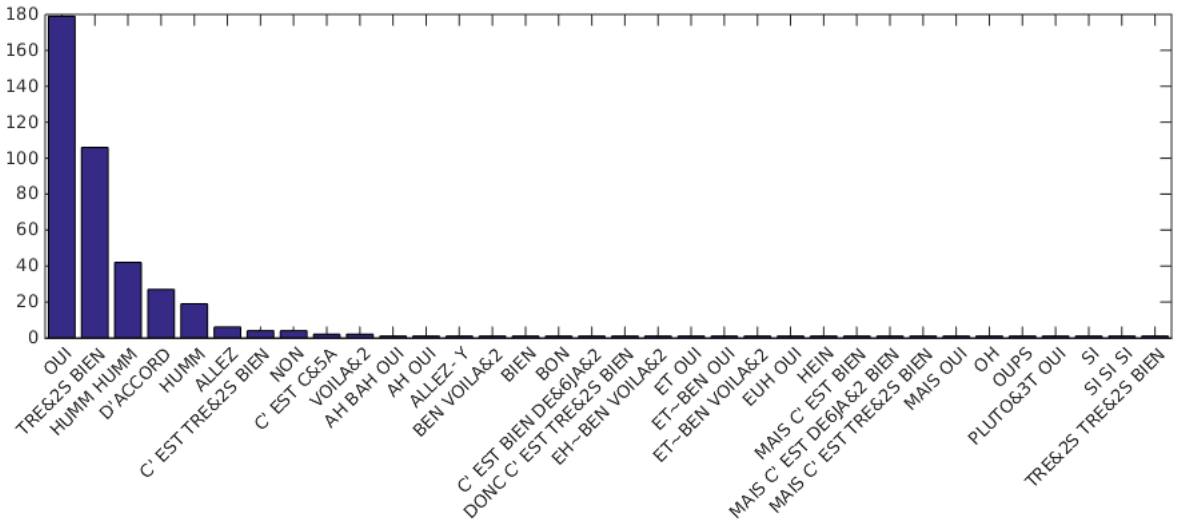


Figure 2.9 – Number of occurrences of the 34 different lexical markers used by the interviewer to encourage the subjects [Bai+16]. This distribution is dominated by 5 items: *oui*, *très bien*, *humm humm*, *d'accord*, *humm*.

2.2.4 Comments

In this scenario, Bailly et al [Bai+16] quantified specifically backchannels such as assessment, incentive, closure of sub-task, optional reply and confirmation. Interactive speech is in fact characterized by a larger number of backchannels such as *oui*, *très bien*, *humm humm*, *d'accord*, *humm*. The interviewer produced around 500 backchannels in the corpus (see Figure 2.9). These backchannels fulfill different functions such as assessing or encouraging interlocutor's delivery of information, replying to doubts, etc. Their lexical contents and prosodic patterns are very important for encouraging cognitive activity. The choice of words and prosodic patterns in fact strongly determine the function of these backchannels.

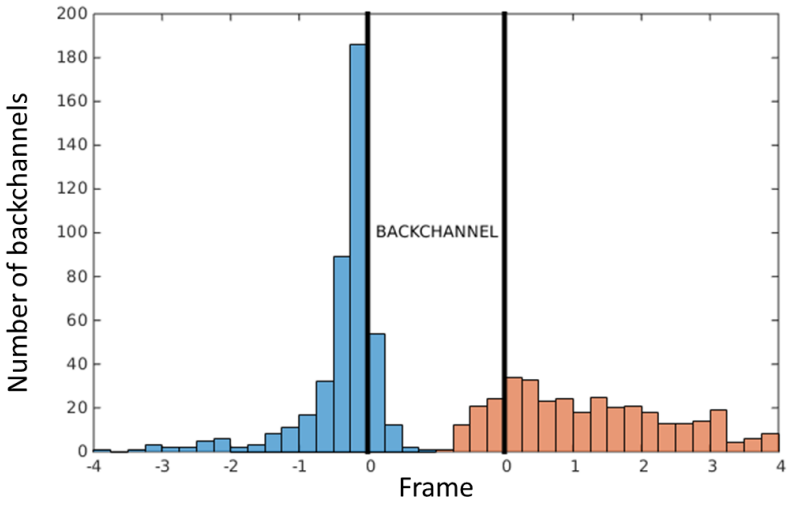


Figure 2.10 – Timing of ends (blue) and beginnings (rose) of interlocutors' speeches surrounding backchannels. The thick vertical bars align all verbal activities to the onsets and offsets of the backchannels (the distance between them is arbitrary and does not reflect the average duration of backchannels). The majority of backchannels are triggered immediately at the end of interlocutors' speeches.

Moreover the majority of these backchannels are timely produced just after one interlocutor's spurt (see Figure 2.10) so that to minimally interrupt his/her turn. In order to generate backchannels, *incremental* interactive models should mainly spot items in the subject's speech in order to instantaneously trigger scoring and interviewer's feedbacks.

2.3 Conclusion and Discussion

In this chapter, we illustrated how our SAR behaviors modeling and evaluation objectives constrained our data acquisition and annotation task, with the example of the two scenarios in which our robot will be involved. The first scenario is a collaborative task - *Put That There* (PTT) scenario where the robot will become an instructor to collaborate with an human

manipulator by giving him/her guidance to move cubes from a reservoir to a chessboard between the human and the robot. In the second scenario, the robot will play the role of a neuro-psychologist to interact with elderly people in a memory test - *Selective Reminding Tests* (RL/RI) scenario. In both scenarios, the robot is expected to generate multimodal interactive behaviors that smoothly engage human subjects. In the PTT scenario, in order to reach effective collaboration with human partner, the task requires the robot to perform precise coordinations of verbal and nonverbal behaviors such as head and gaze movements for joint attention with the manipulator, arm pointing to the cube, etc. That way, if the manipulator cannot capture the instruction from the robot speech (due to the quality of the speech synthesis system), the partner would still be able to finish the task by infering information from other nonverbal behaviors such as gaze, pointing gestures, etc. In the RL/RI scenario, when interacting with elderly person, the robot has to generate high co-occurrences of verbal behaviors with non-verbal behaviors. The task requires the robot interviewer to perform the professional skills exhibited by the psychologist such as engagement, incentive, politeness, etc. in order to encourage the patient trying his/her best to recall learned items (e.g. a lot of backchannels should be used to respond and reward the subject's answers).

In this chapter, we also present how to collect data from human-human interaction and and how to convert the data to useful features that are used in further steps of the learning framework. The features can be used to train the multimodal interactive behaviors for the robot (in chapter 3) as well as to design gesture controllers (in chapter 4) and the speech synthesizer for the robot. In the learning framework, we are interested in modeling multimodal interactive behavioral models at both low- (trajectory of movement) and mid-level (discrete events) that could be used to drive the gesture controllers.

In order to improve HRI quality, the robot should adapt its behaviors to the profile of the current subject, in particular his/her cognitive abilities and physiological capabilities. Therefore, a large set of interactive HHI data should be collected to uncover the impact of these factors as well as social factors such as gender, culture, etc. We then have to find ways to bias interactive behavioral models with parameters related to these *style* features (see current proposals made in the domain of speech synthesis by [Wan+18]) and incrementally estimate these parameters as the interaction unfolds (see the concept of *gesture follower* proposed by Bevilacqua et al [Bev+09]).

Using the large set of HHI data to run HRI also faces the challenge of scaling behaviors from human demonstrators to the robot that has different perception and action abilities. In order to solve a part of this problem, we will further discuss how to collect interactive data by an immersive tele-operation system that enables robot-mediated HHI interaction. Chapter 5 will detail the immersive teleoperation system that is now used in the lab to collect interactive data that is more compatible with the robot's sensorimotor abilities. We believe that this embodiment technique will provide robots with faithful and rich data, as needed in a machine learning framework, with the additional benefit of easing the semi-automatic data annotation.

Chapter 3

Multimodal Interactive Behavioral Models

Human interaction takes place via multimodal behaviors. We use speech, gaze and facial expression as well as head, arm and body movements for communicating with each other in daily life. The purpose of this interaction is not only to transfer task-related specific information in a particular goal [Thó99] but also overt (visible external behaviors, e.g. arm gestures, head motions) and covert mental (internal states, e.g. taking-turn, giving-turn) and physiological states. In order for humans to communicate fluently, many tasks such as speech and scene understanding, dialogue planning, turn-taking, gaze control, arm gestures, etc. should be processed and controlled in parallel. Elementary end actions such as gazing to an agent, nodding or pointing to an object should be coordinated with each other because they often jointly contribute to the encoding of a unique task, such as trying to get the turn [CCD00] – e.g. via gazing, nodding and backchanneling – or attracting attention of the conversational partner to a specific region of interest of the joint space – e.g. via gazing, hand pointing and naming. This multimodal coordination is really important to maintain the conversation. While the"GO" signals of these elementary behaviors are certainly triggered by elementary tasks, speech activity is often used as a baseline timer to which all other modalities coordinate [CCD00].

Note that the multimodal interactive models not only have to capture the coordination pattern between the modalities for a given agent but also inter-personal coordination (relationship between modalities of the human and those of his/her partner). Modeling human-human interactions are in fact complex tasks, because HHI is endowed with multiple modalities that are paced by multi-level perception-action loops [BER08]. An interactive conversation between humans is paced by hierarchical components which have timing relations to each other as shown in Figure 3.1. Particularly, the conversation can be decomposed into turns (1-30 sec) and each turns into multimodal actions, each multimodal action driving motor movements, such as pointing or backchannelling (100-300 msec).

Humans tend to apply social models when interacting with social robots [BSS11] and even apply their social norms (e.g. culture) to decide how to interact with humanoid robots [Hay+07]. In order to achieve effective and natural interaction with humans, ideally, humanoid robots need to perceive, understand and generate interactive multimodal behaviors in order to set the mutual interactive ground that enables the development of the many cognitive and social

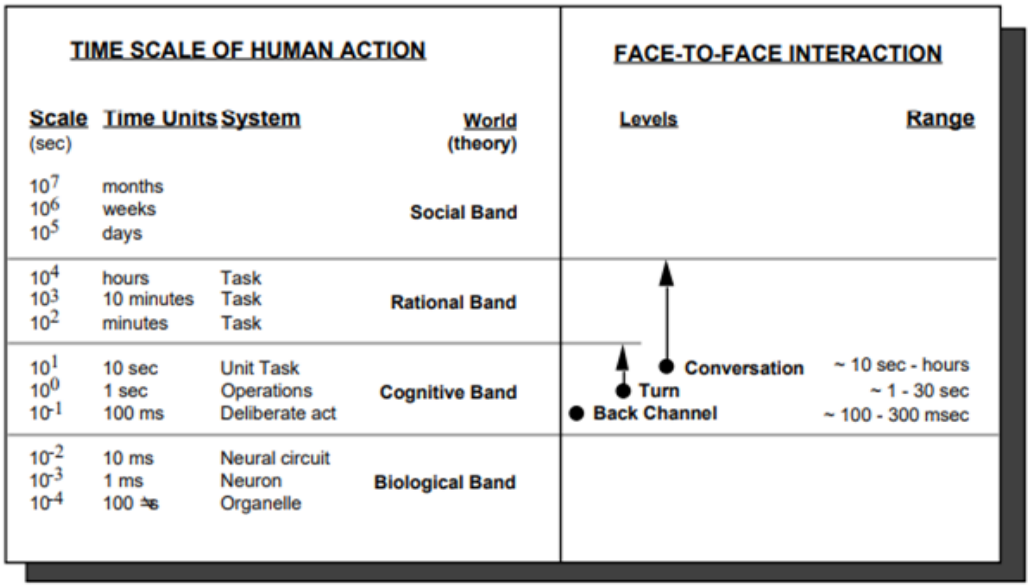


Figure 3.1 – Face-to-face interaction within the time scale of human actions [Thó99]

bounds between rational and emotional agents (see Figure 3.1). Conversational skill is one of the main objectives of our learning framework displayed in Figure 3.2.

Interactive behavioral models are typically built using rule-based methods or statistical approaches such as Hidden Markov Model (HMM), Dynamic Bayesian Network (DBN), etc. In this chapter, we present interactive behavioral models based on recurrent neural networks, namely Long-Short Term Memory (LSTM). We are interested in generating both abstract-level ("GO" signals for elementary movements) and skill-level (trajectory of motions) behaviors:

abstract-level behaviors The behavioral model here triggers elementary actions such as the activation of an eye saccade, a pointing gesture or a back-channel given actions of others and an internal estimation of the state of the conversation. Speech, gaze and gestures of the two subjects involved in the PTT task are here modeled jointly. The results show that the proposed LSTM networks are more effective than the conventional statistical methods in generating appropriate overt actions. We also applied these methods to generate backchannels in the RL/RI task.

skill-level behaviors The behavioral model here maps directly perception to gesture without considering intermediate representations/actions. We generate head motions of the instructor in the PTT task with a cascaded LSTM architecture. This solution can capture the coordination between head motions and the others better than a baseline LSTM model.

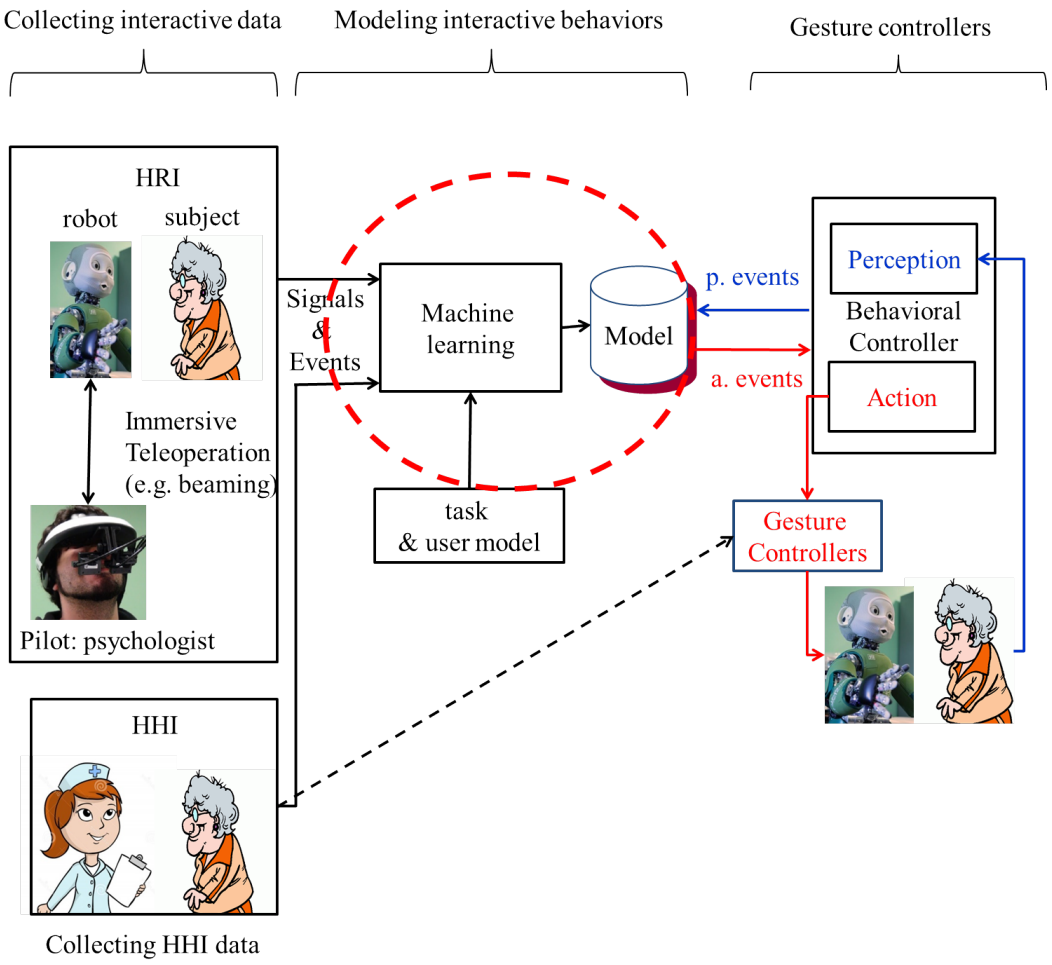


Figure 3.2 – SOMBRERO learning framework: Modeling multimodal interactive behavioral models

3.1 State of the art: modeling multimodal interactive behaviors

This section overviews current approaches for modeling multimodal interactive behaviors. There are classically two main approaches to this challenging issue: rule-based vs. machine learning methods.

3.1.1 Rule-based methods

In rule-based methods, researchers first analyze the recordings of human interactions and try to semi-automatically find lawful patterns in multimodal streams. Computational frameworks are then proposed to operationalize those findings. Such systems usually incorporate set of rules that map perceptual cues to multimodal actions via an intermediate estimation of communicative intentions.

We presents here several examples of rule-based interactive systems.

The BEAT system [CVB01] is quite emblematic of what was developed in the late 90s. It basically augments textual dialog with nonverbal behaviors by enriching the linguistic structure with language tags such as rheme/theme contrasts, objects and actions. The BEAT system extracts linguistic and contextual information from raw input text to control the movements of arms, hands, and face of an avatar as well as the intonation of its voice. A set of rules derived from nonverbal conversational behavior research was used. For example, rules used to control gaze are: "For each THEME: If at beginning of utterance or 70 percents of the time, suggest gazing away from user"; or " For each RHEME: If at end of utterance or 73 percents of the time, suggest gazing towards the user".

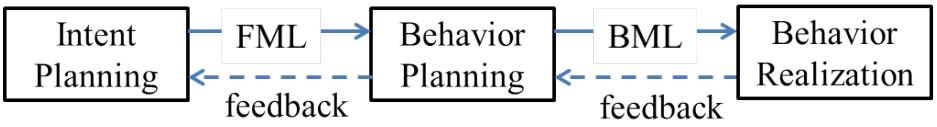


Figure 3.3 – The three stages of SAIBA and the two mediating languages: FML (function markup language) and BML (behaviour markup language). The figure is reproduced from [Kop+06]

```
<bml>
   <speech id="s1" type="application/ssml+xml">
     <text>This is an <mark name="wb3"> example</text>
   </speech>
   <head id="h1" type="NOD" stroke="s1:start"/>
   <gesture id="g1" stroke="s1:wb3" relax="s1:end" type="BEAT">
     <description level="1" type="MURML">...
     </description>
   </gesture>
   <gaze id="z1" target="PERSON1" stroke="g1:stroke-0.1"/>
   <body id="p1" posture="RELAXED" start="after(s1:end)"/>
   <cadia:operate target="SWITCH1" stroke="p1:ready"/>
</bml>
```

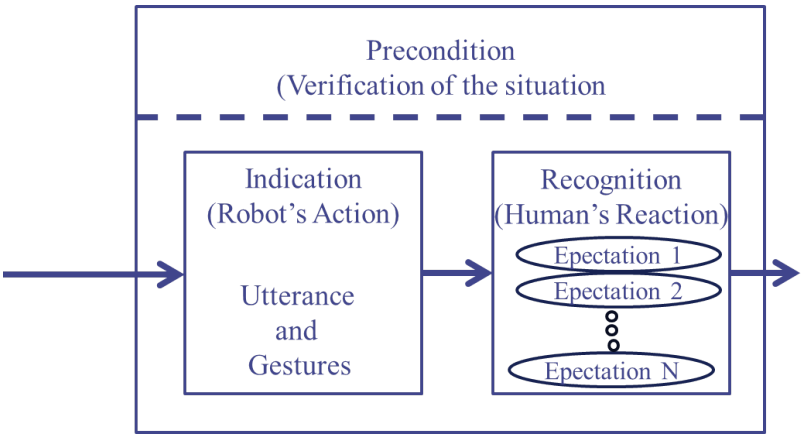
Figure 3.4 – An example of a BML block [Kop+06]

A framework for real-time generating multimodal behaviors is SAIBA (Situation Agent Intention Behavior Animation) [Kop+06]. The framework includes three successive stages: (1) planning communicative intent, (2) planning multimodal realization of the intent, and (3) realization of the planned behaviors as shown in Figure 3.3. There are two mediating XML based languages between the stages: Functional Markup Language (FML) that describes intentions and Behavior Markup Language (BML) that describes nonverbal/verbal behaviors and should be realized by an animated agent. Mutimodal behaviors such as speech, gesture, gaze, body movement, head motion are coordinated in a BML block, which consists of rules. Each behavior is split into six phases which is bound by two of seven *sync-points*: start, ready, stroke-start, stroke-end, relax and end. Behaviors are coordinated by assigning a sync-point

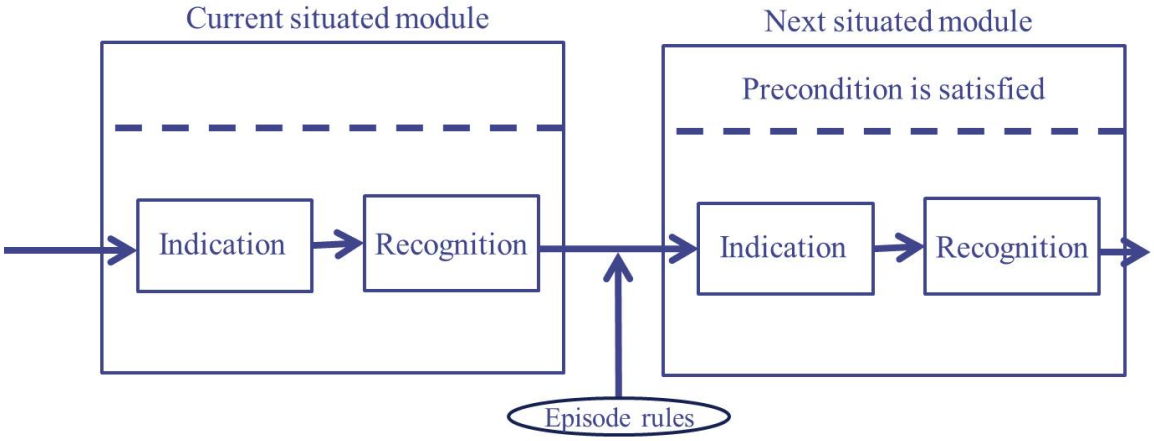
of one behavior to a sync-point of another. Figure 3.4 illustrates an example of a BML block. In this example, a speech tab defines a sentence “This is an example", which is spoken by a text-to-speech system. Head nodding is aligned with the speech's *start* sync-point; and arm gesture is trigged by a *wb3* event which is a new sync-point defined in the speech tab. Based on the SAIBA, Lee and Marsella [LM06] built a Nonverbal Behavior Generator system to generate behaviors according to communicative functions. The system generates nonverbal behaviors such as head movements, facial expressions and body gesture by analyzing syntactic and semantic structure of input text. Particularly, the nonverbal behaviors are assigned with some specific words, phrase or speech acts by rules derived from analyzing a number of video clips.

Thorisson [Thó02] proposed an event-based language where a finite state machine (FSM) describes an interaction scenario as a series of states with pre-conditions and post-actions structured in three hierarchical layers (reactive, process and content). They built a dialogue model, namely *Ymir*, which was used to drive a virtual agent named *Gandalf* in task-oriented dialogues. The architecture includes several modules: perception, decision, as well as knowledge and action scheduler. The perception modules include two types: (1) Unimodal Perceptors that detect important events of single modalities such as prosodic, speech, positional, directional and then (2) Multimodal Integrators that collect all the information from the unimodal ones to come up with a more comprehensive description of user's behavior. The perceptual modules receive and prepare input data to be used as the basis for decisions to act. The knowledge base of the system contains any knowledge that have to do with dialog such as participants, their body parts, etc. The decision modules decide to read mental and world states from perceptive modules and other information from knowledge base module and decide what will be acted. Most of the perceptual and decision modules produce Boolean output (on/off) with the intent to help building larger systems. The decision modules are based on rules and send behavior requests to action modules when preconditions are satisfied and following top-down priority levels: reactive layer (highest), process control layer and content layer. The action modules will manage the behavior requests and execute the behaviors following an any-time algorithm [Dea87] (managing life-span, when activating, deactivating action, etc.).

As another example of rule-based model, Kanda et al [Kan+02] built a tool named *Episode Editor*. The tool is used to drive behaviors of a humanoid robot (a Robovie robot) by building *situated modules* which are orchestrated by *episode rules*. A situated module realizes an action–reaction pair as an interactive and reactive behavior between human and robot in a particular situation. Each situated module (shown in Figure 3.5.a) includes three parts: pre-condition, indication and recognition, which is used to perform certain interactive behaviors such as shaking hands, greeting people, guiding visitors, etc. The pre-condition part verifies if the situated module can be executed or not. Then, if the pre-condition is satisfied, the indication part will generate the robot's action (utterance/ gestures), for example, "lets wave right hand to greet people". After that, the recognition part checks the expected human reaction with regards to the robot's actions generated by the indication so that a human pilot can trigger the most suitable action of the robot. The situated module could be executed consecutively and controlled by episode rules to establish a sequence of situated modules (robot's behaviors) shown in Figure 3.5 (b). One disadvantage of situated modules is their



(a) Situated module



(b) Sequence of situated module controlled by episode rules

Figure 3.5 – Situated modules controlled by episode rules [Kan+02]

limited ability to perform multiple tasks at the same time as well as monitoring complex sequences.

While being quite efficient and easy to deploy for specific interactive tasks, hand-crafted rules have difficulty in taking into account the many factors conditioning the multimodal behaviors (task, personality, social context, emotion, gender, etc.) while maintaining a fine-grained life-like variability.

Another popular approach is based on machine learning techniques which try to find behavior regularities and possibly some of its variability directly from data.

3.1.2 Machine learning methods

The non-verbal behaviors depend on many factors such as gender, cultural, personalities, etc. For example, cultural norms decide for how long or when and in what situations it is appropriate to gaze into another person's eyes. They also exhibit quite some variability, that systematic rules cannot capture. Therefore, with rule-based methods, the implementation of these complex pre-conditions and conditioning variables will be really expensive and time-consuming or just impossible. So, to avoid these difficulties, we need other approaches to more automatically finding useful rules instead of hand-crafted rules. Fortunately, because overt nonverbal behaviors are inherently observable, therefore, supervised machine learning method can be used to train models of communication that can generate robot adequate behaviors. Recently, researchers have begun to study robot behavior models by applying machine learning methods.

As an example, Huang et al [HM14] proposed a learning-based approach using DBN to model human multimodal interactive behaviors – i.e. speech, gaze, and hand gestures – during narration and used this model to generate the multimodal behaviors of a humanoid robot. They defined four streams of observations: a cognitive process (C) that rules how humans coordinate their multimodal behaviors (speech, gaze, gestures). They also assumed that speech (S) further influences gestures (Ge) and gaze (Ga). These causal relations are illustrated in Figure 3.6. Otsuka et al [OSY07] also proposed Dynamic Bayesian Networks (DBN) to estimate addressing and turn taking ("who responds to whom and when?") while using the conversational regime as a latent variable. Similarly, Mihoub et al [MBW15] introduced so-called interaction units – that could be considered as elementary skills or sub-tasks – using Hidden Markov Models (HMM) to generate the gaze of an interlocutor given his own speech activity and the gaze and speech activity of his partner. Then, they improve the model with semi-HMM, which further constrains durations of hidden states. Mihoub et al [Mih+16] then showed that DBN outperform both full- and semi-HMM in predicting co-verbal behaviors in the "Put That There" game.

Actually, few works have been devoted to the modeling of joint behaviors while incredible amount of research have been successfully dealing with recognition of human activities from multimodal behaviors [VNK15]; [Liu+17] and as well as generation of robot behaviors [Nod+14]; [Vog+14].

3.2 Recurrent Neural Network - Long-Short Term Memory

Recently, Deep Neural Networks (DNN) learning gained much success in image processing, speech recognition and speech generation. However, there are few applications of that technique to build models for interactive behaviors. In our work, we applied DNN to incremental sequence-to-sequence mapping where each input frame of perception streams will produce one output frame of action streams.In particular, Long Short Term Memory, a popular method in sequential modeling to generate multimodal interactive behavior such as gaze, arm, head

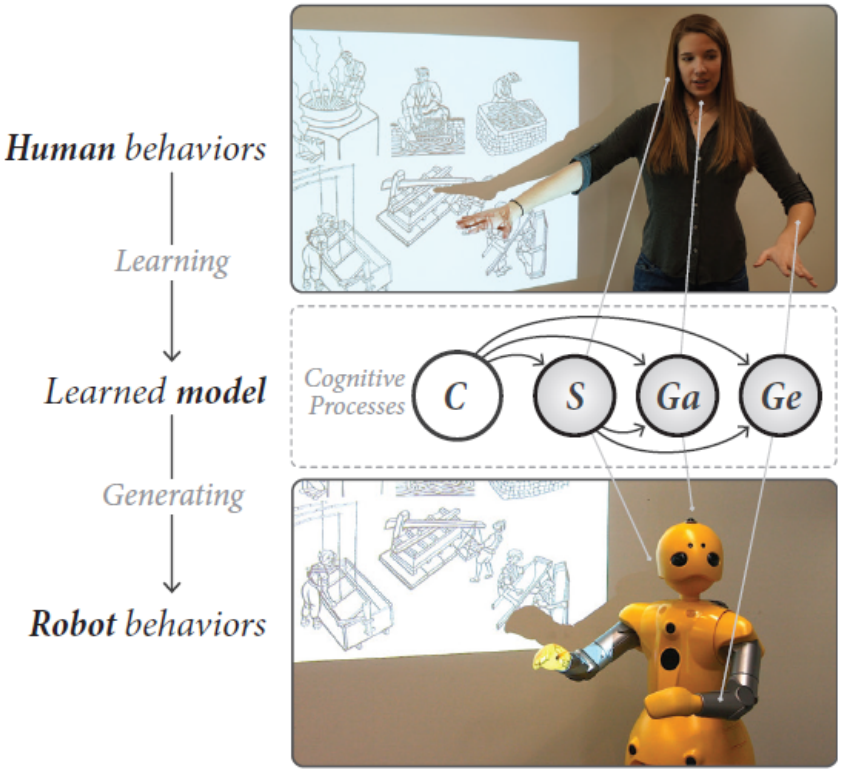


Figure 3.6 – A multimodal interactive behavioral modal using DBN for a humanoid robot in a narration task [HM14]).

movements and backchannels.

3.2.1 Recurrent neural networks

3.2.1.1 Simple Recurrent Neural Network

Recently, recurrent neural networks (RNN) have been applied to sequential data due to its ability to use past information which has been getting through.

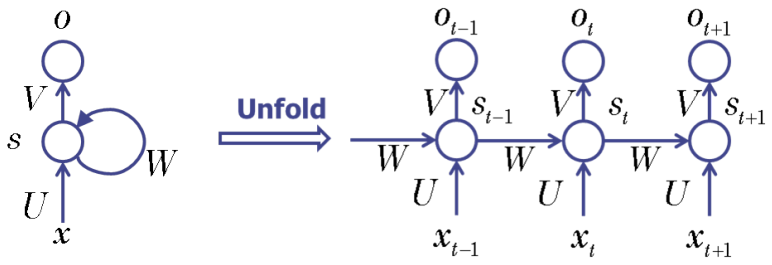


Figure 3.7 – Unfold RNN

The schematic model of a RNN is shown in Figure 3.7, that unfolds the sequential processing into a full network. x_t is input vector of the model at time step t; s_t gathers the hidden states at time step t and could be understood as memory of the network. The hidden state at time step t is calculated based on the current input step x_t and the previous hidden state h_{t-1}, so that it can capture information of all sequence in previous steps, following Eq. 3.1.

$$s_t = f(U \cdot x_t + W \cdot s_{t-1}) \tag{3.1}$$

where, f is a linear or nonlinear (sigmoid, RELU, atan, etc.) activation function.

The output of the network at time step t is o_t calculated from the hidden state s_t by this equation: $o_t = softmax(V \cdot s_t)$. U, V, W are parameters of the networks which should be learned during the training step. Unlike time-delayed neural network, U, V, W are kept the same across time so that the number of parameters of the RNN could be significantly reduced. Also, because of the constraint, the RNN could be expected to ovoid over-fitting.

Ideally, the RNN can model in long-term dependency between hidden states. However, standard RNNs have difficulty in capturing long-term dependencies because of the vanishing problem of fixed feedback i.e. the convergence of geometric series. The vanishing problem occurs during training the neural network using the Back Propagation Through Time (BPTT) method [Wer90]. To look in more detail the vanishing problem, let us calculate the derivative of loss function at time step $t = 3$ with W, which is following chain-rules and illustrated following equation

$$\frac{\partial E_3}{\partial W} = \sum_{k=0}^{3} \frac{\partial E_3}{\partial o_3} \frac{\partial o_3}{\partial s_3} \frac{\partial s_3}{\partial s_k} \frac{\partial s_k}{\partial s_W} \tag{3.2}$$

Because W is used in every step from $t = 0$ to the end of the sequence, back propagation is performed from the current time step t through network all the way to $t = 0$ as illustrated in Figure 3.8. Because each neurons have bounded activations in range of (0,1), the multiplications make the gradient values shrinking exponentially. That means the gradient of a time step far from current one become zero and does no contribute to the processing of the current step. Therefore, the RNN cannot learn long-term dependencies. RNN with gate mechanisms – such as Long-Short Term Memory (described bellow), Gated Recurrent Units (GRU) – have been proposed to solve this vanishing problem.

3.2.1.2 Long-Short Term Memory

Long-short term memory (LSTM) RNN is able to prevent the vanishing problem by adding binary gates to each neuron. These gates determine whether each memory cell should process the available input, use feedback or deliver output (see Figure 3.9). Figure 3.10 illustrates the LSTM unroll in time steps: "o" means the gate is opened – i.e. allows information to pass through the gate – while sign "-" means that the memory cell is closed – preventing information running through. Because the cell is able to close its input gate and thus disable writing to the cell, it may prevent any changes if the cell activity remains unmodified over many time steps, so that longer term dependencies can be learned and preserved.

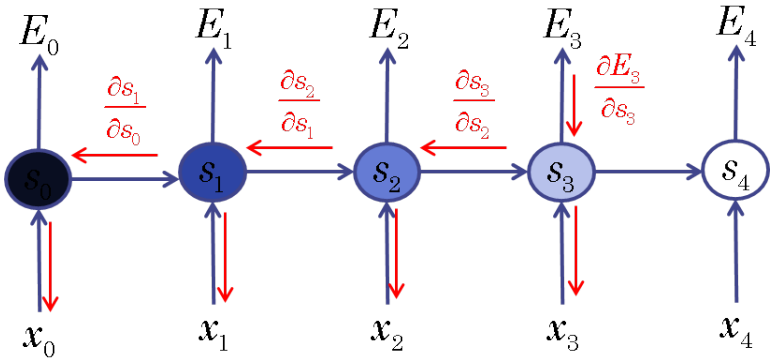


Figure 3.8 – RNN vanishing problem [Den]

The following sets of equations [HS97] illustrate the forward pass of a LSTM block that computes output activations h_t from input activations x_t given the previous internal states c_{t-1} of the units as follows:

$$f_t = \sigma_g(W_f.x_t + U_f.h_{t-1} + b_f) \tag{3.3a}$$
$$i_t = \sigma_g(W_i.x_t + U_i.h_{t-1} + b_i) \tag{3.3b}$$
$$o_t = \sigma_g(W_o.x_t + U_o.h_{t-1} + b_o) \tag{3.3c}$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c.x_t + U_c.h_{t-1} + b_c) \tag{3.3d}$$
$$h_t = o_t \circ \sigma_h(c_t) \tag{3.3e}$$

,where f_t, i_t, o_t are respectively forget, input and output gates; c_t cell states and h_t output of the LSTM block; σ_g, σ_c and σ_h typically are sigmoid, hyperbolic tangent vs. hyperbolic tangent functions.

The architecture shown in Figure 3.11 features a bidirectional recurrent neural network (BiRNN). It consists in combining the processing of the same data sequence in both forward and backward direction performed by two distinct RNN. Their two output layers are then connected to one additional layer that combines the outputs once the whole sequence has been processed. BiRNN has improved the performance in many sequence learning tasks, where the result can be postponed at the end of the sequence [BS14] [GJM13].

3.2.2 Application of RNNs in human interactions

Recently, Recurrent Neural Networks (RNN) have been shown to outperform statistical models in sequence recognition. Gated recurrent units (GRUs) and Long-Short Term Memory (LSTM) cells have been introduced to cope with long-term temporal dependencies. Because of their ability to modulate between short- and long-term dependencies, they are particularly suited for building latent spaces that mediates input-to-output co-variations. Therefore, LSTM becomes state of art of many applications related to sequential data such as statistical language modeling [DMBM15], machine translation [SVL14], and generation of captions from an image or a video [KFF15], etc. Another advantage of LSTM is that it can learn

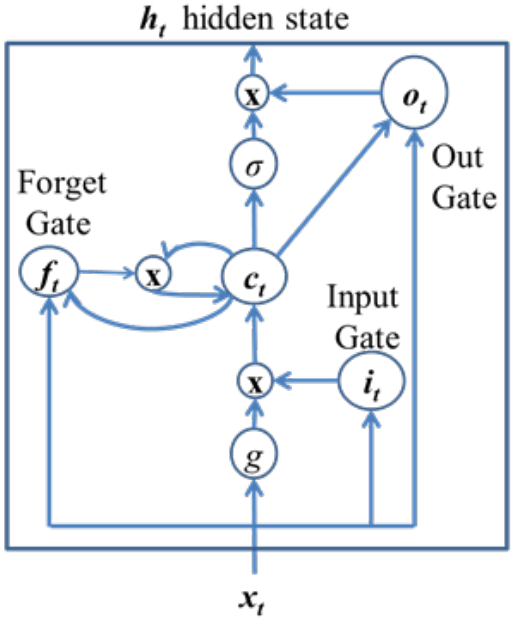


Figure 3.9 – The LSTM unit is a memory block that can be updated, erased or read out according its internal activation c_t and the current input x_t.

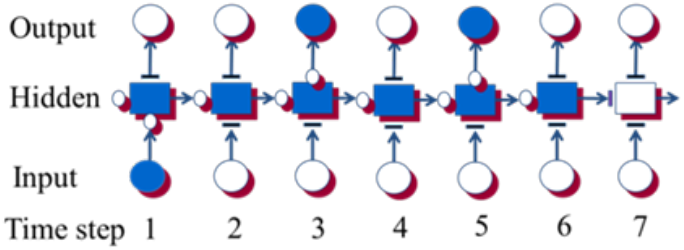


Figure 3.10 – LSTM unroll.

timing intervals between sub-patterns in sequences [GSS02]. Such coordination patterns are particularly crucial to multimodal behaviors such as those involved in natural human-robot interaction.

Most of LSTM-based models have been proposed so far for the recognition of human activities. For example, Ordóñez et al combined Convolution Neural Network (CNN) with LSTM to build a DeepConvLSTM framework which is able to recognize human activities from wearable sensors with minimal pre-preprocessing [OR16]. Furthermore, Tsironi et al also build a CNN-LSTM to learn gestures which have varying duration and complexity [TBW16]. Tian et al [TML15] performed successful emotional recognition in spontaneous dialogs with LSTM.

Fewer works have been devoted to the prediction or generation of interactive behaviors. Alahi et al [Ala+16] used LSTM with social pooling of hidden states which combines the infor-

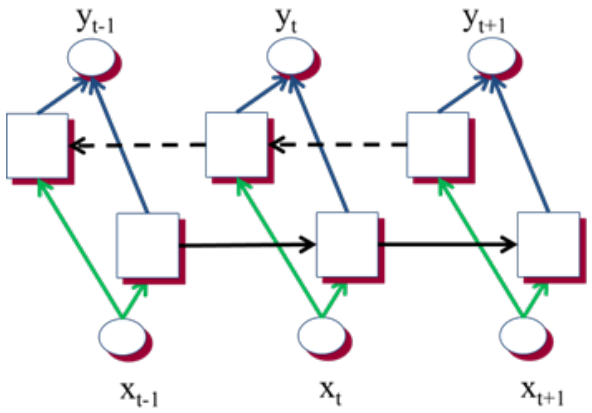


Figure 3.11 – Bi-directional RNN

mation from all neighboring states to predict human trajectories in crowded space. Ravichandar et al [Rav+16] built a promising model of sequential tasks using LSTM in order for a robot to predict what human will do next. LSTM-based conversation models [JH16] have also recently proposed to predict turns in two-party conversations. Schydlo et al [Sch+18] used LSTMs to predict of human intent (multiple and variable length actions) from body pose and gaze cues.

Because of ability to capture long-term dependency of latent variables, we propose to use based on Long-Short Term Memory (LSTM) to model multimodal interactive behavioral models. In the following sections, we present how to build and evaluate the multimodal interactive models to generate discrete variables (gaze, arm, backchannel) and continuous variables (head motions).

3.3 Generating discrete events: Arm and Gaze

In this section, we present multimodal interactive behavioral models using Long-Short Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), that predict gaze and arm gestures in the *Put That There* task. We compare both accuracy and coordination of prediction from our models with those of other methods (using HMM, DBN) proposed by Mihoub et al [MBW15]; [Mih+16] on the same dataset. We tested two versions of each model:

off-line models that perform estimations once the whole sequence has been observe

on-line models that perform estimations incrementally at each time frame.

3.3.1 Modeling techniques

3.3.1.1 Hidden Markov Models

A multimodal interactive model based on HMM was proposed in [MBW15]. In this model, each interactive unit (IU) is modeled by one Discrete Hidden Markov Model (DHMM) that models joint multimodal sensorimotor behaviors via its hidden states. Eq. 3.10 defines the parameters of the DHMM models

$$\lambda_p = (A_p, B_p, \pi_p) \tag{3.4}$$

where $p = 1..P$ is the index of each interaction unit (the number of DHMMs P equals to 6 for the PTT game).

$$O^p = (o_t^p)_{t=1..T} = (SP_t, MP_t)_{t=1..T} \tag{3.5a}$$
$$O^a = (o_t^a)_{t=1..T} = (GT_t, FX_t)_{t=1..T} \tag{3.5b}$$
$$O = (o_t)_{t=1..T} = (o_t^p, o_t^a)_{t=1..T} = (SP_t, MP_t, GT_t, FX_t)_{t=1..T} \tag{3.5c}$$

The observation vectors – T is length of the observation sequence – are separated in two parts: the perceptual streams and the action streams illustrated in Eq. 3.11.

Each DHMM can be trained using Expectation and Maximization (EM) algorithm. The DHMMs were trained with joint streams aligned by IUs. Global transition probabilities between the DHMMs were calculated by a bi-gram model. At training stage, all data streams are available, while in testing only the endogenous verbal stream and exogenous observations are available as shown in Figure 3.12a. After training, two sub-models (a hidden state decoder and an action generator) are thus extracted and used in two steps as shown in Figure 3.12b. Firstly, the hidden state decoder estimates sensorimotor states from perceptual observations only shown in Eq. 3.6. The decoding of sensorimotor state sequence is performed offline by Viterbi alignment and online by a bounded Short-Time Viterbi algorithm with no lookahead.

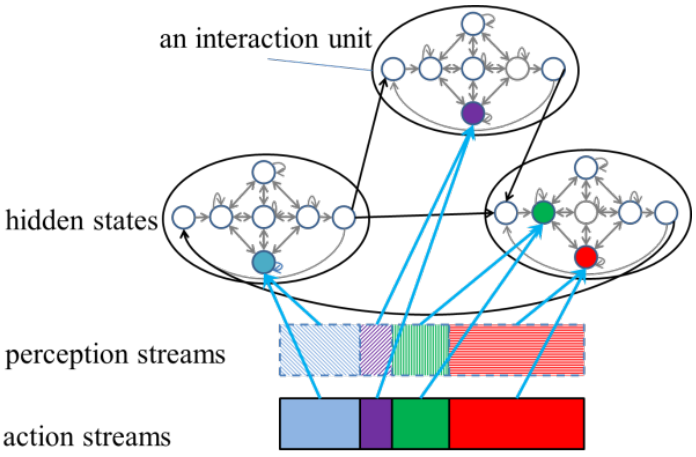
$$S^\star = \underset{S}{\operatorname{argmax}}\, P\left(S|O^P, \lambda\right) \tag{3.6}$$

where S is the sequence of states, $S^\star$ is the optimized sequence estimated from the Viterbi algorithms.

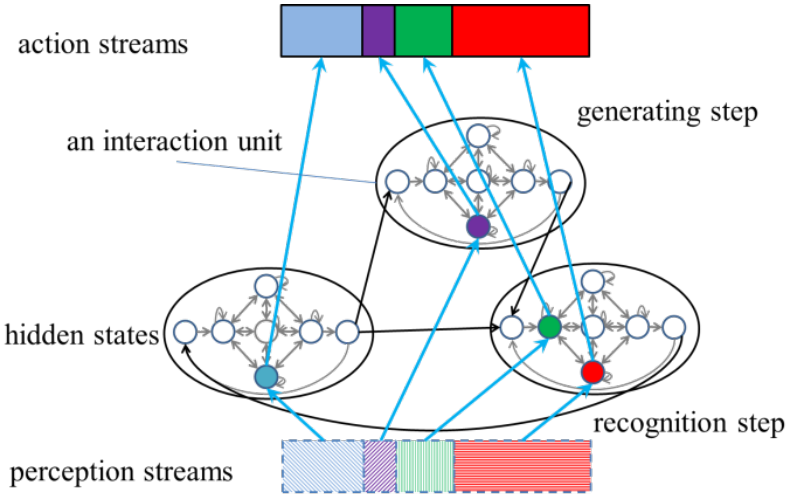
Next, the action generator determines actions from these estimated states as shown in Eq. 3.7.

$$O^A = \underset{S}{\operatorname{argmax}}\, P\left(O^a|S^\star, \lambda\right) \tag{3.7}$$

where, O^A is the stream of actions generated by the generation model.



(a) Training with joint perception and action streams



(b) Estimating hidden states from perception stream and then generating action stream

Figure 3.12 – Schematic of HMM-based multimodal interactive modeling: (a) training (b) generating. Emission is drawn in cyan color and transition probabilities in single DHMM and global HMM are draw in gray and black arrowed lines respectively

The HMM model was implemented with 5 hidden units for each single DHMM using PMTK3 toolkit of Matlab [DM12]. Mihoub et al [MBW15] showed that the results were not improved by using 6 or 7 unit states.

3.3.1.2 Dynamic Bayesian networks

A Dynamic Bayesian network (DBN) is a Bayesian network (BN) with variables linked by temporal dependencies. The network is a probabilistic graphical model that features the probabilistic relationships between random variables via a directed graph (DAG) in which nodes represent random variables and edges present conditional dependencies. A DBN has the ability to deal with uncertainty and to model complex temporal relationship among variables thanks to the intra-slice and inter-slice dependency structures which can be learnt from data by measuring mutual information between children and parent nodes as illustrated in Figure 3.13. In addition, parameters of the DBN model can be also learnt by the Expectation and Maximization (EM) method.

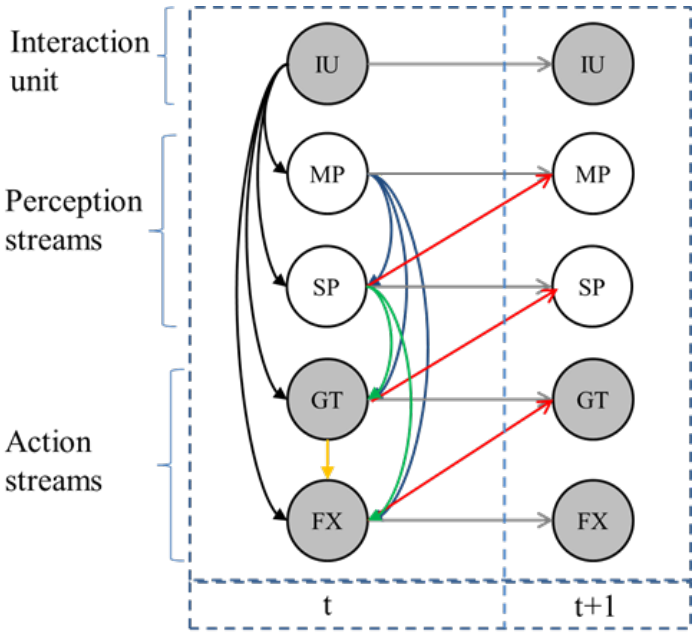


Figure 3.13 – The learned structure of the DBN model: gray circles cue the predicted variables in the inference stage (reproduced from [Mih+16])

The learned DBN model can be used for inference with junction tree algorithm. There are several inference methods to estimate the sequence of actions either on-line or off-line. The *filtering* inference method estimates unobserved nodes $X_t = (IU_t, GT_t, FX_t)$ of the model at time t given the sequence of observed nodes $Y_{1..t} = (SP_{1..t}, MP_{1..t})$ as shown in Eq. 3.8 bellow:

$$X_t^\star = \underset{X_t}{\operatorname{argmax}} P\left(X_t | Y_{1..t}, \lambda\right) \tag{3.8}$$

The *smooth* inference method estimates the action $X_t^\star$ given the whole perception sequence

as given in Eq. 3.9.

$$X_t^\star = \underset{X_t}{\operatorname{argmax}} P\left(X_t | Y_{1..T}, \lambda\right) \tag{3.9}$$

The DBN model was implemented using Bayes Net Toolbox [Mur+01] for inference and training in which the intra-slice structure and inter-slice structure were leant by the K2 [CH92] and REVEAL [LFS98] algorithms, respectively.

3.3.1.3 Long-Short Term Memory

We have built discriminative multimodal interactive models using LSTM so as to improve the sensitivity of internal/latent variables to long-range structural dependencies. LSTM can be trained to directly map perception to action, in particular without necessarily considering a priori knowledge of the underlying structure of the interaction, i.e. the interaction units (IU) introduced by Mihoub et al [MBW15].

We will however show that such a priori knowledge can be beneficial to performance, in particular when training data is limited. A way to implicitely inform the LSTM hidden units about the underlying structure of the interaction is Multi-tasking. Multi-task learning [ZZ14] is meant to (1) implicitly structure the main mapping task by feeding the network with additional objectives and (2) prevent over-fitting with additional and related tasks. We thus applied the multi-task methodology to implicitly structure the prediction of actions (main task) by also predicting interaction units (cognitive states/subtasks). Long-term and short-term processing capabilities of the multi-tasking LSTM are expected to benefit both to high-frequency (i.e. mapping actions) and low-frequency (i.e. recognizing units) tasks.

Figure 3.14 illustrates the training of multimodal interactive behavioral model using multi-tasking RNNs. The main task remains to predict action events (FX and GT) from perceptual events (MP and SP) shown. The secondary task consists in predicting IU. The loss function of LSTM model will thus be the sum of the loss function of IU, GT and FX. Since all variables are discrete with almost identical cardinal, no weighting was performed. Neither did we decrease IU contribution as a function of iterations.

In this research, we build each multi-tasking RNN models with minimal number of hidden layers. The LSTM model has only one LSTM layer with 35 gated units. The BiLSTM model has one forward LSTM layer and one backward LSTM layer with the same number of gated units. The outputs of the two LSTMs are then fed to a time distributed dense layer applied at each time step (i.e. the output layer of the BiLSTM) with soft-max activation functions. The cardinal of the outputs of the forward and backward LSTM as well as the BiLSTM equals the sum of the cardinals of the different classifying tasks. Both of LSTM and BiLSTM model were implemented by using Keras [Cho+15].

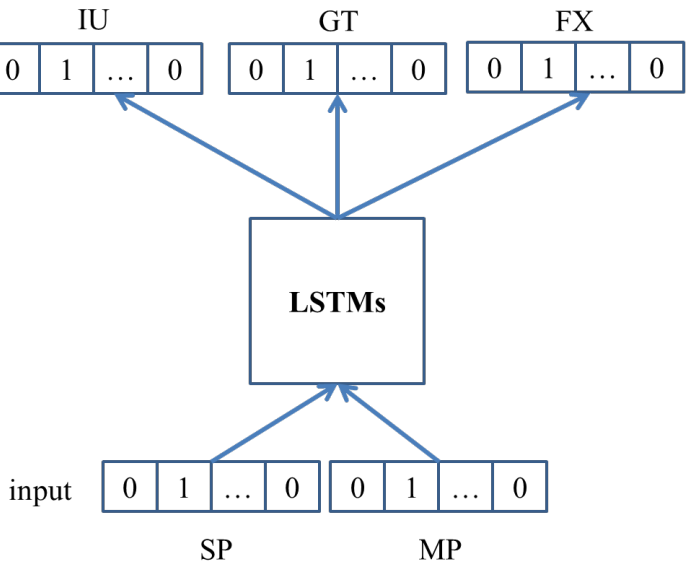


Figure 3.14 – Schematic model of multi-tasking LSTMs. As for DBN, input streams include MP and SP. Identically, output streams are GT and FX. IU is treated as a secondary task for regularization purpose.

3.3.2 Results and Discussion

The LSTM and Bi-LSTM can automatically learn contextual variables from the interaction scenario. In order to compare the efficiency of the methods, actions (FX and GT) generated offline by BiLSTM are first compared with HMM [MBW15] and DBN [Mih+16]. In addition, on-line predictions of the actions by LSTM are also compared with short-term Viterbi decoding of HMM and on-line filter prediction of DBN.

For all models, leave-one-out cross validation is applied to the 30 folded games. Both frame-by-frame comparison and Levenshtein distance estimation [YB07] are performed. We also perform coordination histogram, as proposed by Mihoub et al [Mih+16], in order to compare global coordination patterns between different modalities given synchronous streams of discrete events. A coordination histogram computed for one modality cumulates the delays between each event in this modality and the nearest events observed in the other modalities. We compare the ground-truth coordination histograms with those predicted by the various models.

3.3.2.1 Off-line task

Prediction accuracy Figure 3.15 summarizes the prediction accuracy and Levenshtein distance between ground-truth and models'outputs for off-line prediction tasks. Because of the possibility to a direct conditional dependency between input and output observations, DBN outperform HMM for all features: IU (74% vs. 59%), GT (82% vs. 78%) and FX (61%

vs. 49%). The BiLSTM model surpasses both other methods for IU (79%) and FX (64%) prediction (95% confidence level), respectively, while the accuracy of GT prediction caps at 83%. All prediction accuracy rates are much higher than the empirical chance levels of the tasks, i.e. 21% for IU, 34% for GT and 20% for FX. The same observations apply for the Levenshtein distance. These good results may be explained by the ability of LSTM to learn the complex syntactic organization of the features from the surface structure, notably causal relations that are spanning across IUs.

Figure 3.16 displays chronograms of input and output sequences predicted by the different models. The two first rows show the input sequences from the instructor: speech SP with 5 values (cube, location, reference, none, else) and arm gesture of manipulator MP with 4 values (rest, grasp, manipulate, end). The three final rows superimpose predictions of output streams GT and FX and IU in the different methods to the ground truth. Most onsets of predicted events by BiLSTM for the output streams are close to onsets observed in the ground truth, while onsets predicted by HMM are generally the most distant ones. This is confirmed by evaluating coordination histograms (see next).

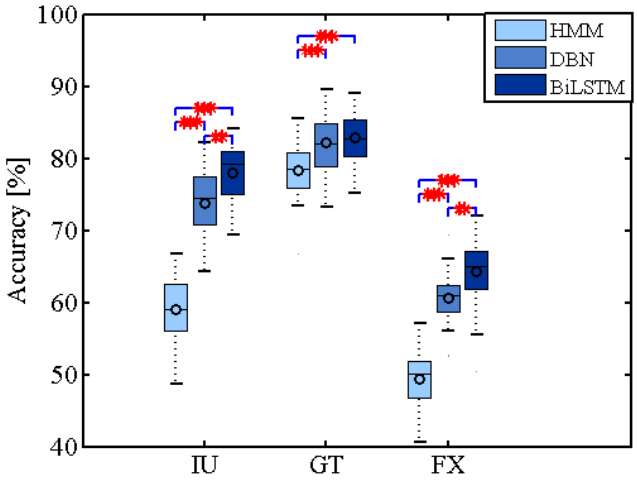
Table 3.1 – Chi squared distances between the coordination histograms of ground truth vs. those of the different off-line models. Note that degrees of freedom ($df < 10$) depend on the distribution of delays in the different percentiles. Since events are sampled at 25Hz, the minimum bin is 40ms.

Stream	*HMM*	*DBN*	*Bi-LSTM*	*df*
SP	1054	78	**72**	**8**
GT	783	375	**112**	6
FX	1327	199	**92**	**8**

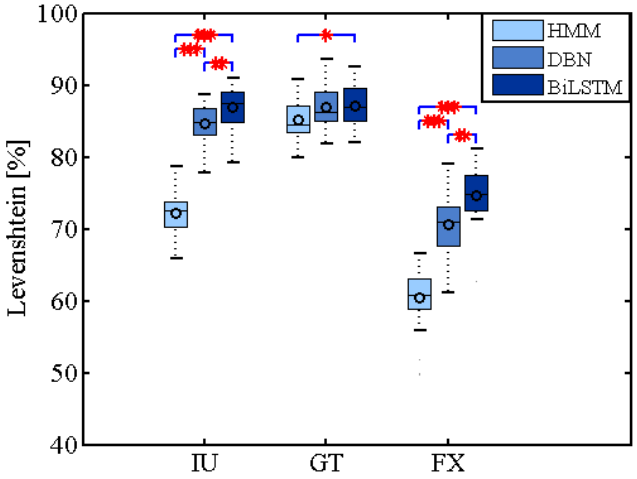
Coordinate Histogram Coordination histograms give a global picture of the micro-coordination patterns between each modality and the other ones. These histograms proposed by Mihoub et al [Mih+16] basically collect the delays between events in one modality and the closest ones in the others 3.17. Figure 3.18 shows coordination histograms for ground truth (first row), BiLSTM (second row), DBN (third row) and HMM (final row) corresponding to SP (first column), GT (second column) and FX (last column). Pearson's chi-squared ($\chi 2$) distances between the histograms of the ground truth and the different models are calculated and shown in Table 3.1. Note that cue-specific bins are computed as 10-quantiles of the distribution of events collected by all systems. All histograms significantly differ from each other ($p < 1e^{-3}$) except DBN and Bi-LSTM for SP. The smallest $\chi 2$ distances are those of BiLSTM, which demonstrates that the BiLSTM generates the most faithful behavioral coordination patterns.

3.3.3 On-line tasks

One of the main challenges of the multimodal interactive behavioral model is to on-line feed the gesture controllers of one humanoid robot in face-to-face interaction with a human partner.

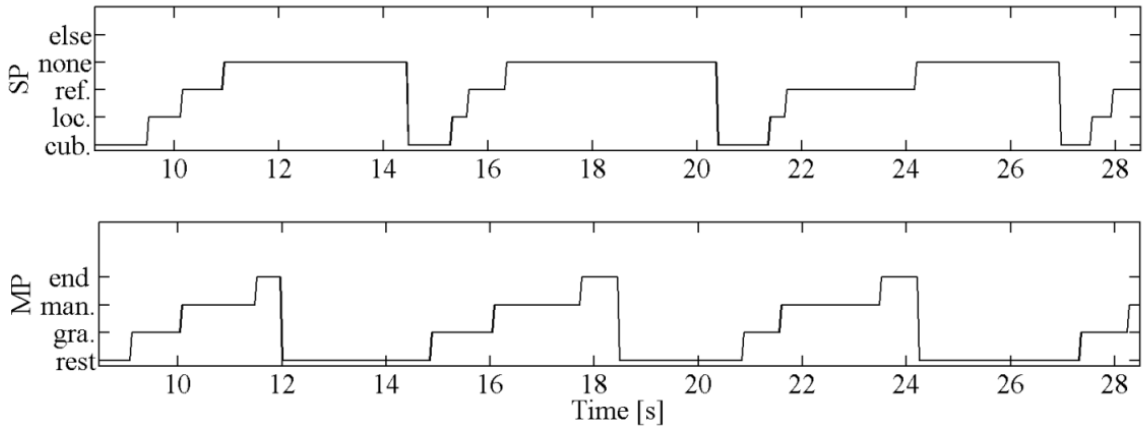


(a) raw F-score

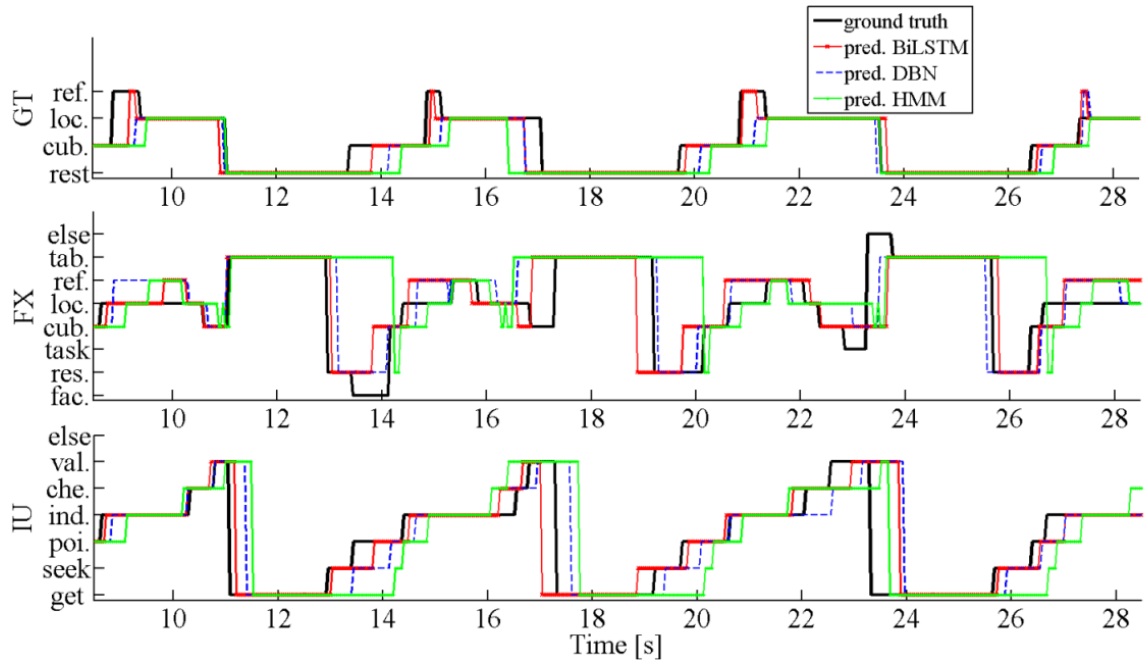


(b) F-score with relaxed alignment

Figure 3.15 – Offline generation: comparing performance of the joint estimation of the 3 different streams (IU, GT, FX) with the methods HMM, DBN vs. BiLSTM. (a) raw F-score, (b) F-score with relaxed alignment. The number of stars above the links between scores cue significant F-probability of Tukey post-hoc tests $'\star\star\star'$ with $p < 1e^{-3}$, $'\star\star'$ with $p < 1e^{-2}$, $'\star'$ with $p < 0.05$). For each box, the internal line gives the mean value of the score while the circle gives its median value.



(a) Input sequences



(b) Output sequences

Figure 3.16 – Input and output sequences: (a) the two top inputs MP and SP. (b) superposition of ground truth and output streams (GT, FX) and IU estimated by the different methods proposed in the paper

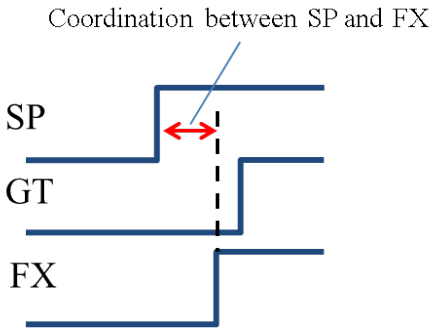


Figure 3.17 – Computing coordinate histogram corresponding to SP by cumulating delays between each SP event and adjacent events in the other two streams GT and FX.

For this purpose, the model's output should be computed incrementally as the input sequence unveils.

Table 3.2 – Chi-squared distances between the coordination histograms of ground truth vs. those of the different on-line models

Stream	*HMM*	*DBN*	*Bi-LSTM*	*df*
SP	1114	1167	**253**	6
GT	1225	1004	**252**	4
FX	749	402	56	7

The comparison of exact-rate prediction and Levenshtein distances for all of the methods in the on-line prediction tasks are respectively shown in Figure 3.19(a) and Figure 3.19(b). Similarly to the off-line results, with Levenshtein estimation, DBN significantly (with 95% confident level) outperforms HMM for both IU (69.64% vs 67.64%) and FX (64.31% vs 60.97%) predictions. While the GT prediction of LSTM is almost the same as the others (84.72% for LSTM, 84.87% for DBN, 83.85% for HMM), LSTM surpasses the other methods for the prediction of IU and FX at respectively 82.93% and 70.72%.

Similarly to Table 3.1, Table 3.2 gives the χ-squared distances between coordination histograms of the ground-truth and predictions of the three methods with the different cues. All histograms significantly differ from each other ($p < 1e^{-3}$) except HMM and DBN for SP. Again, the smallest distances are those of LSTM method. These results show the effectiveness of LSTM in online prediction of faithful multimodal streams which are properly coordinated with each other.

3.3.4 Discussion

The LSTM behavioral model benefits from extracting contextual information from data, instead of being limited to the boundaries of the hidden states of HMM or the immediate previous frames of the DBN dependency graph. We explored several ways to introduce latent variables in the DBN structure, notably by using HMM states as additional latent variables. This does

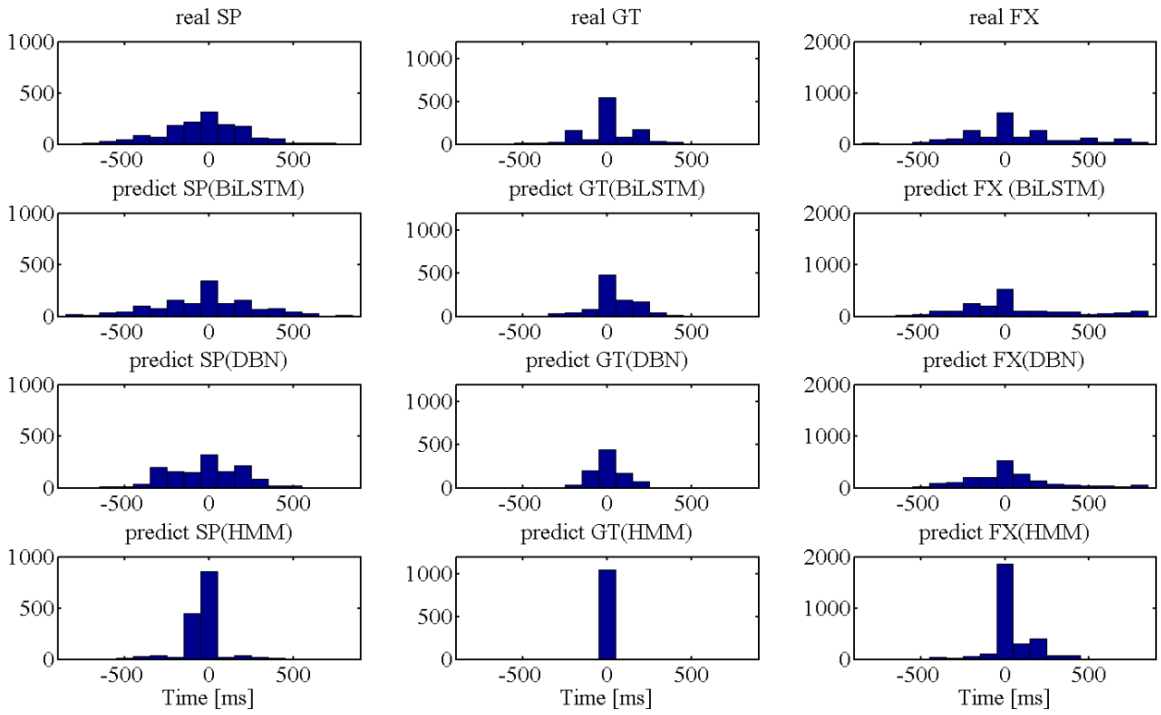
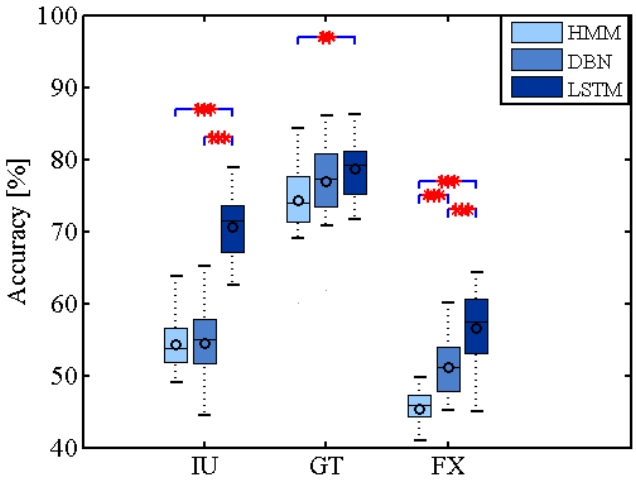


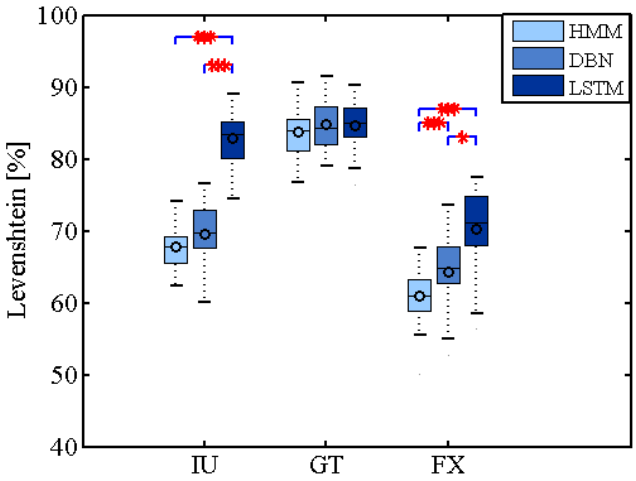
Figure 3.18 – Comparing ground truth coordination histograms (top) with those computed with streams predicted by different offline methods, from top to bottom BiLSTM, DBN and HMM respectively. (a) speech coordination with gesture and gaze (b) gesture coordination with speech and gaze (c) gaze coordination with speech and gesture.

not improve DBN performance in any way. In contrast, LSTM behavioral model has the possibility to collect contextual information far away in the past history. Contextual information may in fact span large lags. As an example, Richardson et al [RDS08] have notably shown that a listener will most likely be looking at an object 2s after his/her interlocutor has been paying attention to. Mihoub et al [MBW15] have effectively shown that adding one frame at around 2 seconds before the current input as contextual information optimally boosted HMM performance for gaze prediction from speech activity. Coordination histograms show that ground truth intermodal coordination does not exhibit fixed delays between events but a rather complex cue-dependent distribution. LSTM has the capacity to modulate memory span according to the current input and the progress of the interaction without unnecessarily increasing the input window.

Note also that our task involves a sequence of elementary interactive skills (our IUs) with low complexity. We expect the ability of LSTM to implicitly stack features to ease the carry-over of information when the task complexity increases.



(a) raw F-score



(b) F-score with relaxed alignment

Figure 3.19 – Performance of the different methods for the on-line prediction tasks. Same conventions as for Figure 3.15

3.4 Generating discrete variables from continous estimations: Backchannels

Backchannelling (BC) makes conversation smooth and natural. These short feedbacks uttered by the listener signal that he/she is understanding and paying attention to the speaker. Especially, when talking with patients, psychologists or coaches produce numerous backchannels to encourage them and reward their responses.

Verbal backchannels uttered by conversational agents improve the quality of human-computer interaction. A key research problem for building immersive virtual humans and robots is learning to predict timing of these backchannels [Mor10]. Most of previous research used statistical models to predict and generate backchannels from produced and perceived audiovisual features. We here present a backchannel predictor based on a Long-Short Term Memory (LSTM) recurrent neural network. Training and test data have been collected during interviews conducted by a professional neuropsychologist. The proposed system aims at complementing a spoken dialog system with automatic generation of backchannels during active listening of the speaker's interventions: the challenge is to predict backchannels utterered by the interviewer given parts of speech uttered by both the subject and the interviewer. F1-measures of backchannels *opportunities* are computed to compare the proposed predictor with a baseline model using Conditional Random Fields (CRF). Subjective ratings of the effective generations of verbal backchannels are also performed. The LSTM model outperforms the state-of-the-art model both in terms of prediction accuracy, alignment with speech turns and subjective ratings by third parties.

3.4.1 State of the art

Two main approaches have been proposed for BC generation: rule-based vs. data-driven methods.

Rule-based Ward and Tsukahara [WT00] fine-tuned a pitch-pause model by analyzing English and Japanese conversations. The model generated backchannels after detecting backchannel opportunities, cued by a downward pitch slope lasting at least 110 ms followed by a pause of at least 700 ms.

A rule-based method was also proposed by Truong et al [TPH10] using pitch and pause information. A backchannel is triggered when a pause exceeds a certain length and is led by a falling or rising pitch. Cathcart et al [CCK03] built a Pause Duration Model which decides if backchannelling occurs at a *Transition Relevance Place (TRP)* where the listener may take over or not. They argue that most of TRPs contain backchannels.

Data-driven More recently, data-based methods have been proposed. Nishimura et al [NKN07] used the decision tree method to generate dialog system's responses and their timings based on

prosodic features and response preparation status of user. Morency et al [MKG10] evaluated several probabilistic methods such as HMM and CRF. Several input features such as eye gaze and prosody are combined to make prediction.

Ruede et al [Rue+17] proposed to detect BC opportunities using LSTM. The model gets a modest F1-score (0.37) with prosodic features as input. The score raised to 0.39 when adding linguistic information such as word embeddings. In another work, Maier et al [MHS17] also built a LSTM model which uses acoustic and linguistic features to predict end-of-turn events incrementally for situated spoken dialogue systems. Skantze [Ska17] also used LSTM with both voice activity and pitch features to predict whether a turn-shift will occur.

Our work will focus on generating BC for the humanoid robot to perform the RI/RL scenario. Interactive data was described section 2.

3.4.2 Interactive data: Train and validation data

The interactive data includes conversations of an interviewer with 4 different subjects. Leave-one-out validation is performed to optimize generation thresholds: the training set includes data from 3 subjects while the remaining one is used for validating.

For BC detection and prediction, acoustic features such as pitch slopes and pause lengths are mostly used. However, by accumulating the timing relation between interlocutors' utterance around backchannels, Bailly et al [Bai+16] revealed that in the context of the RL/RI scenario, the majority of backchannels are trigged in order to both confirm correct responses as well as foster further retrieval items. Therefore, selected features should directly consider the cognitive activities of each person in the interaction such as introducing the sub-tasks, questioning, correcting answers, etc. In particular, there are two input streams including:

- SI speech of the interviewer with 5 discrete values: introduction the task, pause , give question, listen, feedback. In prediction task, the *feedback* value is treated as *listen*.
- SS speech of subjects with 3 discrete values: speaking, listen, good answer.

3.4.3 Methodologies

Similar to [MKG10], BC are generated in two steps:

prediction: where our model predicts the probability of a backchannel to occur, frame-by-frame

effective generation: backchannels are generated by thresholding this time-varying probability distribution.

In the interactive scenario, the prediction of a continuous backchanneling probability is performed every 40ms on the basis of the speech activity of both speakers. Actual backchannels

are then triggered when this probability exceeds an optimal threshold (computed on the basis of Precision-Recall curves as described in the following).

The prediction task here consists in estimating BC as a function of SI and SS activations. We compare performances of two prediction techniques: Conditional Random Field (CRF) vs. LSTM.

3.4.3.1 Conditional Random Field

A Conditional Random Field (CRF) is a statistical modeling method, more specifically a discriminative undirected probabilistic graphical model [LMP01]. The discriminative model make conditional independence assumptions among output sequences $\mathbf{y}$ but do not assume conditional independence among input sequences $\mathbf{x}$ [SM+12]. CRF can have much simpler structure than a joint HMM model [Mih+16] and thus requires a smaller amount of data. This is rather appreciable for interactive scenario where the data are scarce and difficult to collect.

The conditional probability backchannel predictor $\mathbf{y} = BC_{1..T} = [y_1, y_2, ..., y_T]$ given observed sequences $\mathbf{x} = [SI, SS]_{1..T} = [x_1, x_2, ..., x_T]$ is given by Eq. 3.10:

$$p(\mathbf{y}|\mathbf{x}) = \frac{\prod_{t=1}^{T} exp(\sum_{t=1}^{K} \theta_i f_i(y_{t-1}, y_t, x_t))}{Z(\mathbf{x})} \tag{3.10}$$

where, $Z(\mathbf{x})$ is an input-dependent normalization function:

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{t=1}^{T} exp(\sum_{t=1}^{K} \theta_i f_i(y_{t-1}, y_t, x_t)) \tag{3.11}$$

and $f_i(y_{t-1}, y_t, x_t))$ is feature function, which illustrates the conditional dependency of y_t on x_t and can be understood as a partial *likelihood* of each possible y_t given x_t, θ_i are the parameters to be learned. Based on the equations, HMM can be considered as a particular case of CRF where the feature functions f_i are indicator functions and θ_i are constants, which are used to model state transitions.

In the first-order CRF used by Morency et al [MKG10], the current BC only depends on the local neighborhood of the input features and the preceding BC estimation. The first-order CRF is similar to a first-order HMM which modeling the dependence of the current frame with the previous one, but the former tries to directly model the conditional distribution $p(\mathbf{y}|\mathbf{x})$ while the latter explicitly attempts to model a joint probability distribution $p(\mathbf{y}, \mathbf{x})$. Therefore, the generative HMM results in poor accuracy, which is the metric of interest [SP03], especially in case of data where true positives are rare. On the contrary, the CRF attempts to differentiate between the two situations [MKG10].

3.4.3.2 Input window

We built two other prediction models by supplementing the CRF and LSTM models with delayed input, i.e. the current frame $[SI, SS]_t$ together with $[SI, SS]_{t-W}$ frames shown in Figure 3.20. More precisely, the current frame will be concatenated with previous W frames to enlarge the input of the models. The respective models are referenced as CRF_W and $LSTM_W$. We varied the time shift W so that to get the maximum increase of the F1-score. The accuracy effectively increases with the number of frames. However, when W becomes large, over-fitting occurs that decreases the accuracy. The optimal empirical value for W equals to 5 for CRF vs. 6 for LSTM (i.e. 200 ms vs. 240 ms).

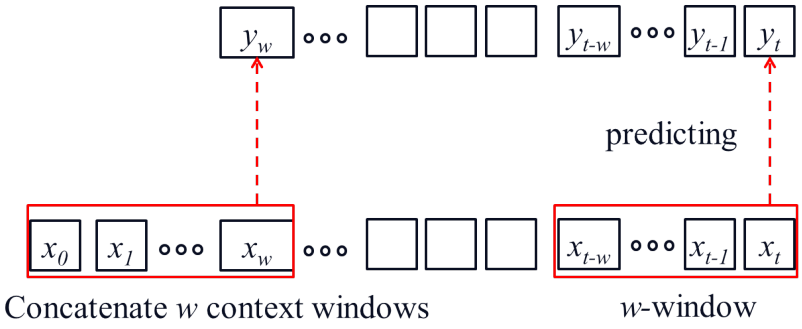


Figure 3.20 – Inputs concatenated w-context windows

3.4.3.3 Implementation

The CRF method is implemented by using HCRF toolkit [HLY06]. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is used to train parameters of the CRF models. The LSTM method is implemented using the Keras framework with the Theano back-end [Cho+15]. The LSTM layer is set with the standard activation functions *tanh* for input, output and forget gates; and *hard sigmoid* activation functions for recurrent steps. We used a single LSTM layer with 30 units and a softmax output layer is used for two classes (yes/no). A Cross-Entropy loss as a cost function and the Adam optimizer [KB14] are used to train the model. Grid-search is implemented to find the optimal number of epochs. Because of the limited amount of interactive data, leave-one-out validation is used to select the best number of epochs (n=75) that gives minimal number of errors in the validation set.

3.4.4 Backchannel prediction and generation

In the prediction task, the SI stream considers ground-truth BC signals as a listening activity. The prediction task outputs BC opportunities as continuous probabilities. These probabilities are then thresholded to actually generate BCs for an effective incremental human-robot interaction: a BC is generated when the probability exceeds an optimal threshold. Note that the BC generator then remains silent for 3 seconds.

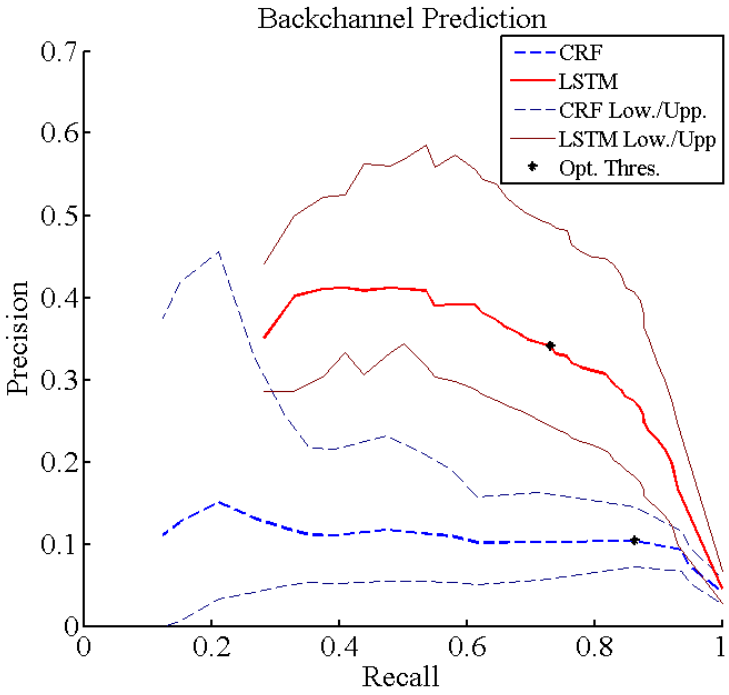


Figure 3.21 – Precision-Recall curve of backchannel prediction with CRF and LSTM

Figure 3.21 features precision-recall curves for CRF vs. LSTM in the prediction task (horizontal axis is recall while vertical axis is precision). The precision-recall curve of CRF method (displayed in blue) lay rather below that of LSTM method (displayed in red). This confirms LSTM predicts back-channeling opportunities more precisely than CRF. The two black star points give the respective optimal thresholds for which the distance of (precision, recall) is nearest from (1,1). Table 3.3 shows the precision, recall and F1-measure at these optimal thresholds. The LSTM method can predict BC (F1-score = 0.434) far better than CRF model (F1-score = 0.185).

Table 3.3 – Precision, recall and F1 score of the two methods in BC prediction task.

Methods	*Precision*	*Recall*	*F1-measure*
CRF (baseline)	0.104	0.863	0.185
LSTM ($W = 0$)	0.342	0.731	**0.439**
CRF ($W = 5$)	0.317	0.762	0.434
LSTM ($W = 6$)	0.342	0.731	**0.486**

Figure 3.22 displays the BC predictions for CRF and LSTM methods, respectively. In many cases, CRF over-predicts BC: BC opportunities are sometimes generated before the subject's answers or predicted during effective silent listening frames. The LSTM seems to better capture the relevance of BC in this task as well as their relative short durations, i.e. mainly signaling correct answers or too large periods of verbal inactivity.

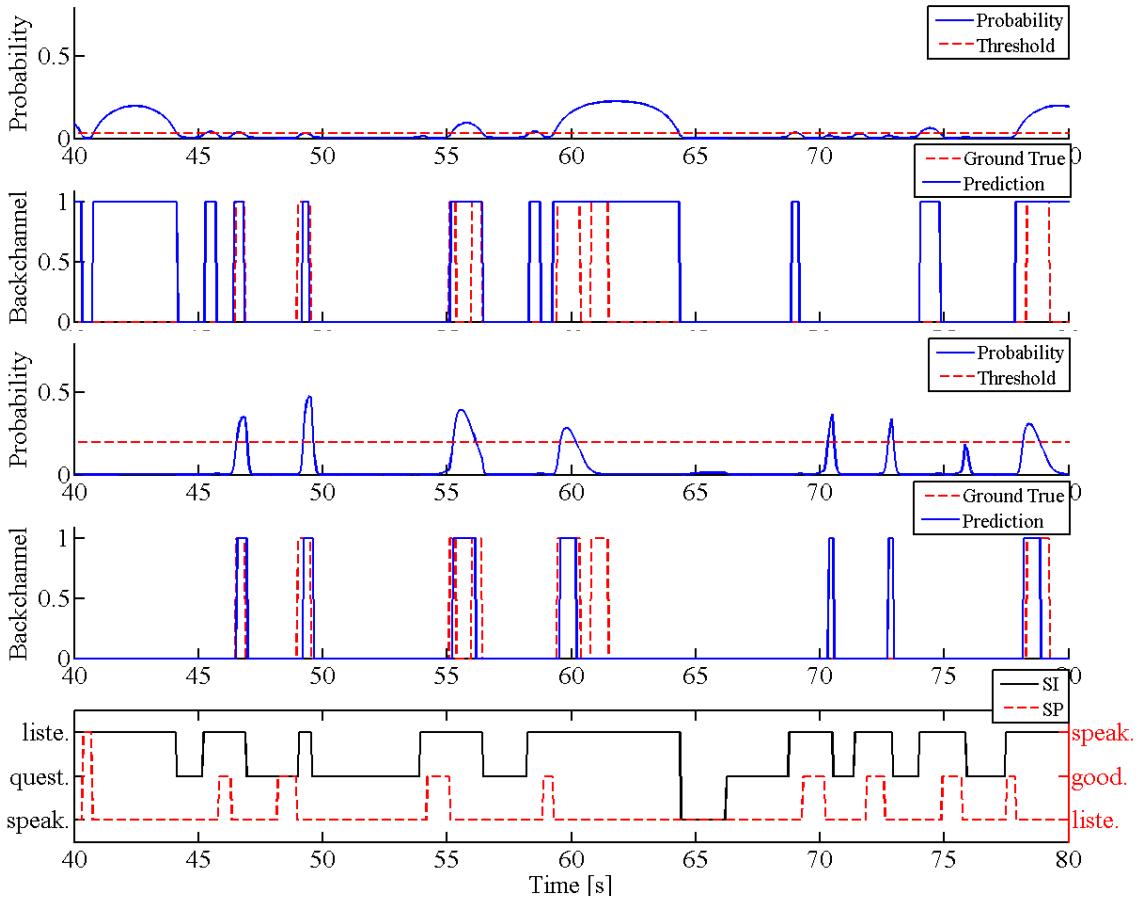


Figure 3.22 – Backchannel prediction and generation by CRF (2 top captions) vs. LSTM (2 next captions) from speech activities of both speakers (bottom caption). Effective generation is performed by optimally thresholding backchannel prediction. Underlying BC activity is considered as "listening" activity in the input interviewer's speech stream.

Figure 3.23 displays the number of BC generated by the different methods together with the original ones for the different subjects (1-4). While the number of BCs generated by the CRF baseline model strongly differs from the ground-truth, the CRF_W and $LSTM_W$ get closer to the empirical distribution. They however still over generate BC for interviewees 1 and 2. Are these extra BC however acceptable? Are they relevant and acceptable BC opportunities that the interviewer did not catch? We performed a subjective evaluation to resolve this issue.

3.4.5 Subjective Evaluation

We evaluated if the BC events generated by the predictor are acceptable in a simulated interaction. We effectively generate BCs and insert synthetic verbal BC at the places predicted by the different models. We then ask third parties to rate the acceptability of these talk spurts. Note that this process is also performed for the ground truth, for the sake of fair comparison: original BC are thus erased and replaced by synthetic verbal BC triggered at the same ground

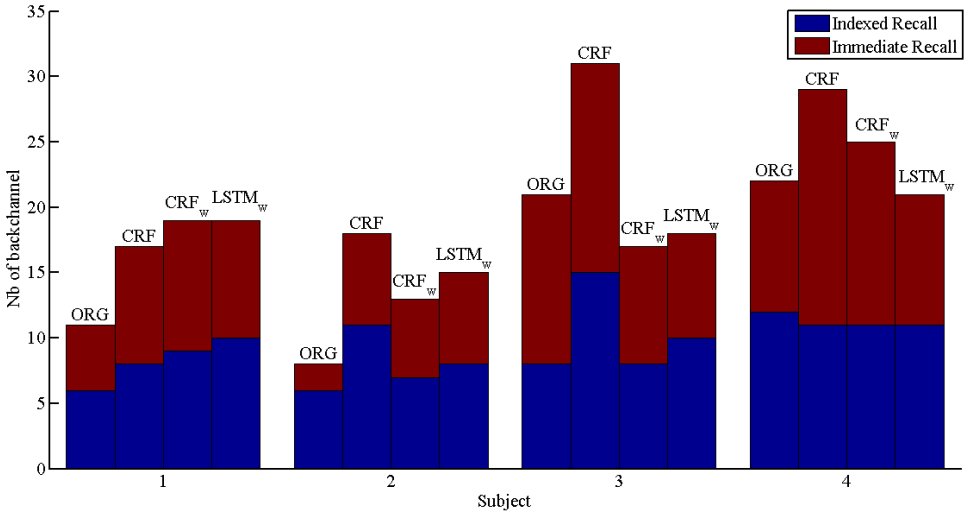


Figure 3.23 – Number of backchannels generated by different methods (ground truth, CRF, CRF_W, $LSTM_W$) for four subjects and original one during the two steps of the learning phase (immediate vs. indexed recall).

truth onsets.

3.4.5.1 Lexical Choice

The BC generators trigger BC events, but they do not now predict the verbal content nor the prosodic patterns of the BCs. During our interviews, the interviewer used 34 different lexical markers to encourage the subject [Bai+16]. Bailly et al showed that 5 lexical markers dominate the distribution: "oui" (yes), "très bien" (excellent) and "d'accord" (okay) together with basic continuers "humm-humm" and "hum". In this evaluation, BC are chosen randomly among the five lexicons following their empirical distribution shown in Figure 2.9. We here pick natural BC whatever their prosodic patterns.

3.4.5.2 Relevance of BC generation

Kok et al [KH12] surveyed methods for evaluating BC generations. For example, Mohri et al [MPR08] compared two methods by asking participants to rate short human-robot interactions with a five-level Likert scale. Similarly, Huang et al [HMG10] evaluated BC generation via a para-social consensus sampling method. In their work, participants saw videos of a virtual agent giving feedback to a speaker by using head nods. Multi-criteria questions are then used to evaluate rapport, believability, wrong head nods & missed opportunities.

Following [MKG10], we asked subjects to rate the relevance of BC generation after listening short speech excerpts centered on one unique BC in context, i.e. starting 2.5sec before the BC

and ending 0.5sec after it. The evaluation was performed by crowd sourcing: we released a website [1] where participants where asked to rate BC generated by either models in a standard five-level Likert item ("very bad", "bad", "why not", "good", and "very good").

In order to focus on the analysis of differences between models, we decided to exclude BCs generated almost at the same positions as the ground truth (i.e. whose onsets differ by less than 0.5sec). This procedure eliminates about half of the generated BCs. We end up with 76, 30 and 36 BC for CRF, $LSTM_W$ and CRF_W respectively. 66 original ORG BC were also considered.

We got ratings from 31 participants (20 males, 11 females, age = 32.5+/-9). Each participant heard 40 BC, 10 for each of ORG, CRF CRF_W vs. $LSTM_W$. Remember that the BC of original interviews were also replaced by synthetic BC.

Figure 3.24 gives the evaluation results. All distributions are statistically different (R package "ordinal", $p < 1e^{-12}$). Only subject 4 significantly deviates from this behavior. CRF gets the worst score while $LSTM_W$ got the highest score. CRF_W was rated better than the baseline but still generates a significant amount of "bad" BCs. The under-performance of ORG may be due to the randomly selection of synthetic BC or the non conservative behavior of the interviewer who sometimes interrupts the interviewee.

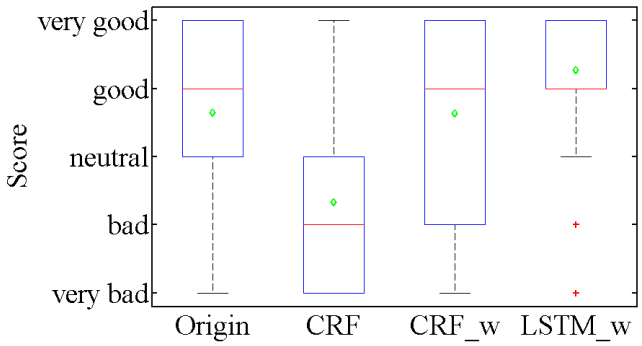


Figure 3.24 – Subjective evaluation results

3.4.6 Discussion

The strategy consisting in augmenting the input frame with w previous frames (called full-concatenated models) actually improves the prediction results with an optimal value $w = 5$ for CRF and $w = 6$ for LSTM. Similar results can be obtained by concatenating the current frame with only one past frame at distance w as illustrated in Figure 3.25 (called fix-concatenated models). Table 3.4 illustrates the optimal distance frames $w = 5$ of two methods (CRF and LSTM) using this fix-concatenated window strategy but have significant lower performance than the LSTM full-concatenated model.

[1] http://www.gipsa-lab.fr/ duccanh.nguyen/bch_evaluation_short/

Although without window contexts, the LSTM (no windows) model still generates backchannels better than CRF ($w = 5$). This confirms again that LSTM model can learn automatically the relationship between the frames through its hidden states.

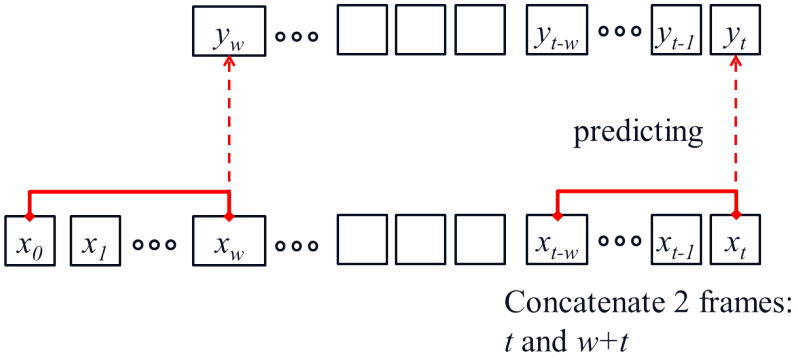


Figure 3.25 – Inputs with concatenated two frames t and t+w

Table 3.4 – F1 score of the two methods with inputs concatenating current frame and only a past frame with distance w.

Methods	*F1-measure*
CRF (baseline)	0.185
LSTM (no windows)	0.439
CRF ($w = 5$)	0.433
LSTM ($w = 5$)	0.443

For now, our works just focus on when backchannels are generated, but in fact, in order to apply to real robot, which lexicon of backchannels generated are also really important. Bailly et al [Bai+16] showed that lexical choices depend on not only their respective frequencies but also conform to syntactic constraints (e.g. there is a high probability of *humm* than *humm humm* after a backchannel *d'accord*). In fact, lexicon of backchannels should be chosen according to their functions. In this task, there are five main functions of backchannels: assessment, incentive, closure of subtask, optional reply and confirmation. The joint prediction of "When and What to backchannel" by statistical methods would require more training material.

3.4.7 Conclusion

In this section, we compare two methods (CRF vs. LSTM) in both prediction and generation tasks using data from a dialog-oriented scenario. In both cases, LSTM models outperform CRF models. LSTM seems to better capture the relevance and timing of BC in the dialog scenario, where the interviewer usually uses BC to encourage the elderly people to be confident in answering questions.

The results from LSTM will be used to generate on-line BCs for a humanoid robot [NBE16] that autonomously performs the neuro-psychological test. An on-line situated evaluation [NBE17]

will be performed to subjectively evaluate the BC generator. Another issue is evidenced in fig. 3.23: $LSTM_W$ still generate too many BCs, in particular for interviewees 1 and 2, with no memory impairments. BC generation should therefore be modulated by high-level settings such as incremental estimation of success rate.

Note that verbal contents and prosodic patterns of BCs should also be appropriately chosen [Bai+16] according to the sub-task and the previously generated ones. The joint prediction of "When and What to backchannel" by statistical methods would require more training material.

3.5 Generating continuous variables: Head Motions

Head motion contributes to multiple functions such as visual attention, emotional display as well as back-channeling and is influenced by multiple social, physiological and cognitive factors.

In this section, we investigate how to generate continuous head motion in the context of the collaborative scenario where head motion contributes to co-verbal as well as non-verbal behaviors. We show that in this scenario, the fundamental frequency of speech (F0 feature) is a poor predictor of head motion, while the gaze significantly contributes to the head motion generation. We propose a cascaded Long-Short Term Memory (LSTM) model that first estimates the gaze from speech content and hand gestures performed by the partner. This estimation is further used as one of the inputs for the generation of the head motion. The results show that the proposed method outperforms a single-task model with the same inputs.

We analyze here head motion data of a human subject involved in a face-to-face cooperative interactive game (see section 2) that both elicits verbal communication and visual attention. We challenge the problem of generating continuous head movements from speech activities and gestures of both partners. We will show how the exploitation of the main causal relations between speech, gestures, gaze and head motion into the modeling architecture benefits to both prediction accuracy and coordinative structures.

3.5.1 State of the art

3.5.1.1 Head motion, gaze and speech

The study of human eye-head coordination during orienting movements to targets has a long history [GV87]. For example, when the head is free to move, the amplitude of the eye saccade is a function of head velocity (e.g. the faster head movements, the smaller the eye movements). Those coordination is influenced by numerous factors including the nature of the target, its position in the field of vision and with respect to the previous fixation, etc. Freedman et al [Fre01] studied the coordinations between eye and head movements in single gaze shift and showed that, in saccades with moving freely head, gaze amplitude is linearly related to

head amplitude. In addition, when the amplitude of the head motions are increased, eye velocity declines and eye movement duration increases. Fang et al [Fan+15] studied eye-head coordination during visual search and revealed that there are multiple saccades during a single head movement and the peak of the distribution of eye fixations is biased toward head orientation. Nakashima et al [NS14] suggested that visual search performance is best when eyes and head are oriented in the same direction. This could suggest a relationship between head and eye movements of the instructor in PTT scenario while most of the time his tasks are finding target positions (e.g. cube in reservoir space, target location, reference cube, etc.). For a conversational system, if the directions of the head and of the gaze are too different, the conversational system can produce an unnatural rotation of eyes [Bev+07].

Head motion also contribute to active listening: it complements binaural cues [BBA13] and has been shown to enhance automatic source diarization and localization [MMB15]. Head motion is also important to acknowledge or replace verbal back-channels (e.g., nodding for acknowledging or shaking for signaling doubt), but also for many aspects of human communication. Munhall et al [Mun+04] showed that vision of head motion improves speech perception. Graf et al [Gra+02] demonstrated that the timings of head motion and the prosodic structure of the text are consistent and suggest that head motion is useful to segment the spoken content. Yehia et al [YKVB00] notably evidenced high correlation between head motion, eyebrow movements and the fundamental frequency (F0) of speech. Head motion also provides useful information about the mood of the speaker [Bus+07].

3.5.1.2 Generating head motions

Rule-based systems are common methods to monitor human interactions. For example, Liu et al [Liu+12] proposed to generate head pan by analyzing utterance structure and identifying backchannels, while head tilt was depending on phrase length. Lee and Marsella [LM06] proposed Generator system that associates multimodal patterns with given communication functions. The system can generate head motions and other non-verbal behaviors from surface texts. They defined several primitive head movements: nods, shakes, head moved to the side, head tilt, pulled back, pulled down. Then, they created rules of the movements matching with their functions (e.g. co-occurring with words) as shown in table 3.5. For an example, a surface text is given: *"I do, I do. I'm looking forward to that but I can't rest until I get this work done."* Rules applied to the text are: **affirmation** rule from *I do* and *I'm*; and **negation** rule from *can't* (contrast rule applied from *but* is overridden by the negation rule. Therefore, head motions will be generated as follow: head nods on *I do*, *I do* and *I'm looking forward*; head shakes on *I can't rest*.

Machine learning techniques have been proposed to generate head motions. For example, Busso et al [Bus+07] proposed to use Hidden Markov Models (HMM) to drive head motion from prosodic features. Ben Youssef et al [BYSB13] used articulatory features to drive head motion synthesis. Another HMM-based framework to generate body movement from prosody was proposed by Levin [LTK09]. Ding et al [DPA13] also trained an HMM to generate head and eyebrow movements. Mariooryad et al [MB12] further explore dynamic Bayesian networks

Table 3.5 – Head motions generated by rules [LM06]. The priority of each rule is marked by the numbers in the parenthesis

Functions	*Head motions*	*Rules*
(1) **Interjection**	Head nod, shake, or tilt	co-occurring with these words: *yes, no, well*
(1) **Negation**	Head shakes	throughout the whole sentence or phrase these words occur: *no, not, nothing, cannot*
(2) **Affirmation**	Head nods	throughout the whole sentence or phrase these words occur: *yes , yeah, I do, I am, We have, We do, You have, true, OK*
(3) **Obligation**	Head nod	once co-occurring with these words: *have to, need to, ought to*
(3) **Assumption**	Head nod	throughout the sentence or phrase when occurring these words *I guess, I suppose, I think*
(4) **Contrast**	Head moved to the side (lateral movement) and brow raise	co-occurring with these words *but, however*
etc.	...	...

(DBN) for coupling speech with head and eyebrow movements. More recently Sadoughi et al [SB17] introduced latent variables to consider speaker intentions.

Recently, Recurrent Neural Networks (RNN) have been shown to outperform statistical models in sequence recognition and generation. Gated recurrent units (GRUs) and Long-Short Term Memory (LSTM) cells have been introduced to cope with long-term temporal dependencies. Few works using LSTM have been performed to model human machine interaction. For example, Alahi et al [Ala+16] used LSTM with social pooling of hidden states which combine the information from all neighboring states to predict human trajectories in crowded space. Haag et al [HS16] proposed Bidirectional LSTM with stacked Bottleneck feature to improve the quality of head motion generation.

In this section, we present a multimodal behavioral model to generate the head motion of an instructor during a collaborative task with a manipulator (the interactive data are described in section 2.1). We proposed a cascaded multi-task learning method, where gaze prediction is considered as an intermediary task for further improving head motion generation. The results can be used partially to drive multimodal interactive behaviors of a humanoid robot.

3.5.2 Multimodal interactive behavioral models for generating head motions

In previous research, Mihoub [Mih+16] built multimodal behavioral models which are able to generate GT and FX given input streams SP and MP. At training stage, all of discrete streams (IU, GT, FX, SP and MP) are available, while in generating stage, only SP and MP are observed.

In this work, we investigate interactive models to generate continuous variables - head motions of the instructor head (H1, H2, H3) with the same observed input SP and MP.

3.5.2.1 Analyzing data

Canonical Correlation Analysis (CCA) is often used to measure the interdependence between two sets of sequential data with different or equal feature dimensions [DFC00].

Using CCA, we computed the correlations between each modality (IU, GT, FX, SP, MP and F0) and head motions (H1, H2, H3). Figure 3.26 displays the mean correlations of this analysis. For all of angles, the highest correlations are with IU and FX. The pitch angle (H1) exhibits the highest mean correlation (0.7) while the others are about 0.5. This can be explained by the fact that FX on face, source's cube space, tablet are well separated in pitch direction while the roll and yaw (H2, H3) are actually not separated into different azimuthal regions. The least correlated feature with head motions is F0! Therefore – for this specific collaborative task – F0 is a minor predictor of head motion. This is expected since speech chunks partly refers to movable regions of interest in the visual scene that are intrinsically referred via nonverbal signals such as gaze.

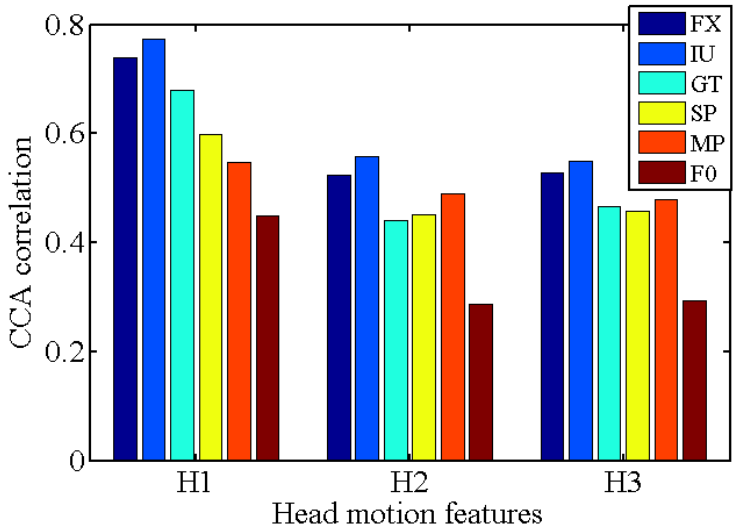


Figure 3.26 – Correlation of CCAs between each of H1, H2, H3 and FX, IU, GT, SP, MP and F0

We compare here the performance of mainly three different models:

Baseline. The baseline model for generating head motion uses one LSTM layer with linear activations to generate directly H1,H2,H3 as shown in Figure 3.27(a). This model uses the same inputs than the DBN proposed by Mihoub [Mih+16], i.e. the observed variables (SP and MP).

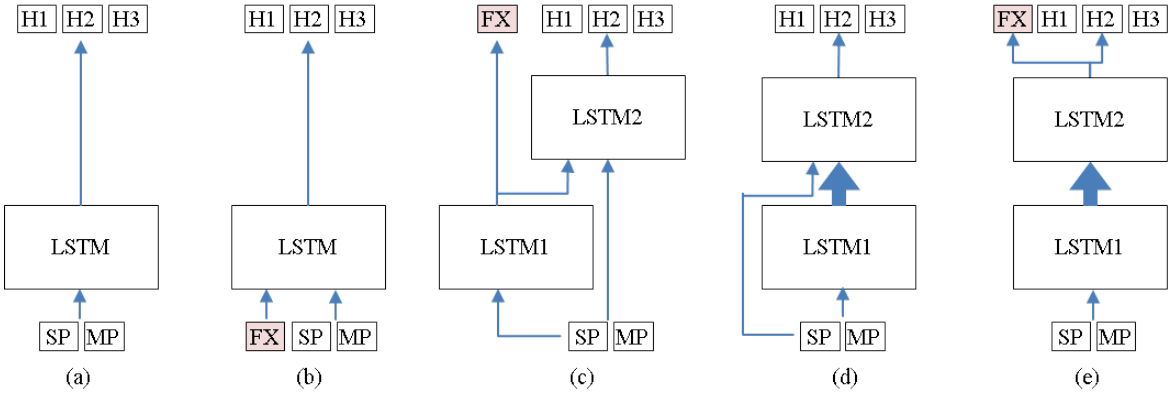


Figure 3.27 – Single vs. Multi-task models: (a) **Baseline** model with inputs (SP,MP); (b) The **control** model with additional FX modality; (c) **Cascaded** model that combines the prediction of FX by LSTM1 with the prediction of head motions by LSTM2 using combined input; (d) **Cascaded single output** model without the intervening FX-prediction task; (e) **Cascaded multiple outputs** model predicting both FX and Hs by LSTM2

Control. Based on CCA analysis results (IU and FX have higher correlation with H1), the control model uses FX as an additional input feature as illustrated in Figure 3.27(b). Our generation models will compete with this control model that is informed by the FX ground truth. Since the correlation of FX with H1 is the highest comparing with H2 and H3, the model is expected to improve significantly the H1 generation quality.

Cascaded. In practice, neither FX nor IU can be used as input feature to train and test data since they are not always available and need to be inferred from observed data such as SP and MP [Mih+16]. The incremental estimation of IU is rather difficult with no look-ahead of observations. On the other end, FX are much more likely to be estimated on-line. We thus propose to use a multitask learning, in which FX is generated by a first LSTM layer, called LSTM1 in Figure 3.27(c). The output of this model is then aggregated with the original input and fed into a second LSTM layer, called LSTM2. This multitask model – with discrete FX and continuous H objectives – is trained in two steps: LSTM1 and LSTM2 are first trained separately and fine-tuning is further performed on the multitask model with both outputs: FX and (H1, H2, H3).

Two other models also have been considered, for fairness:

Cascaded single output. This single-output cascaded model has the same structure as the cascaded model but without the intervening FX-prediction task shown in Figure 3.27(d).

Cascaded multiple output. Including two LSTMs stacked to each other and predicting both FX and Hs illustrated in Figure 3.27(e).

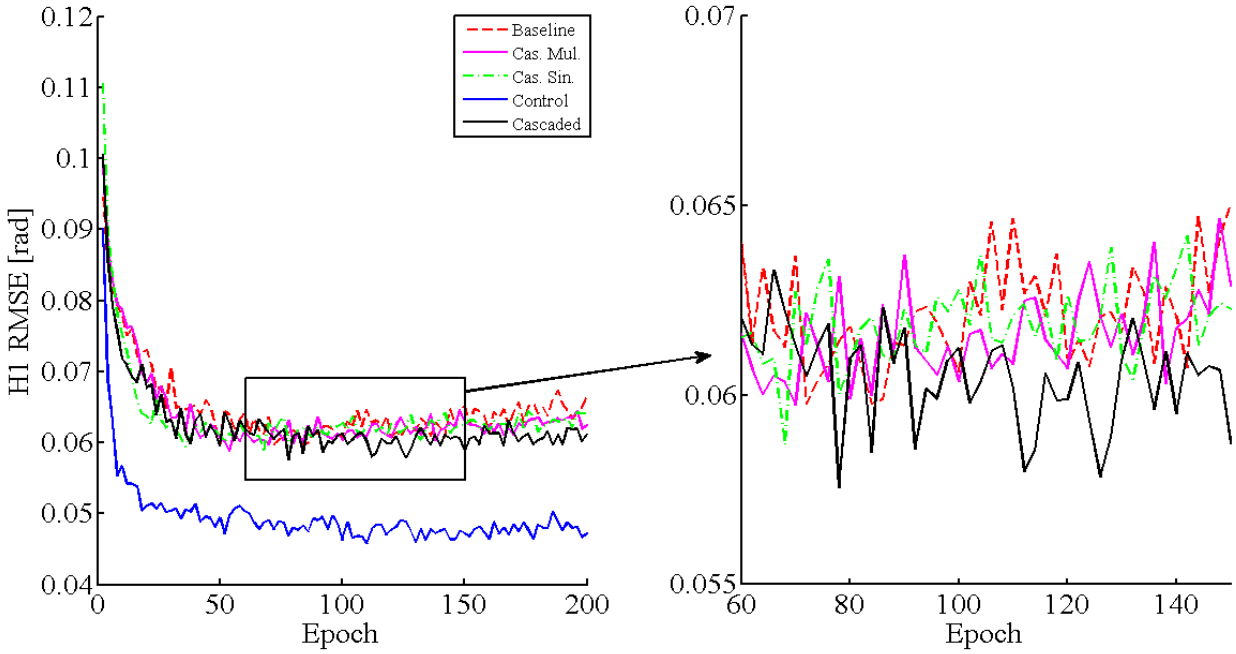


Figure 3.28 – Average H1 RMSE at different epochs corresponding to the different cascaded models.

3.5.2.2 Results

To compare the performance of each model, leave-one-out cross-validation was performed in which 9 interaction sequences were used to train while the remaining one is used for testing. All models use a total of 80 LSTM neurons. Each layer of the cascaded models (LSTM1 and LSTM2) has thus 40 LSTM neurons. Pre-training for LSTM1 & LSTM2 and fine tuning are both performed with 50 iterations. All models are implemented on Keras with Theano back-end.

Figure 3.28 displays root mean square error (RMSE) of H1 as a function of number of epochs and model. The control model clearly outperforms the others at epoch 110 with a RMSE of 0.045 rad. While the *Cascaded* model is able to handle over-fitting and get a minimum RMSE of 0.057 rad at epoch 76, other methods tend to over-fit sooner.

Figure 3.29.a displays a chronogram of ground truth vs. predicted H1. As expected, the *Control* model generates the most faithful movements notably in the vicinity of FX events (see around 7.0 sec). In contrast, the H1 generated by baseline model (driven by the sole SP & MP events) generates delayed head motion. The head motion generated by the cascaded model is close to the one generated by the control model, notably respecting coordination with gaze shifts.

Table 3.6 gives root mean square errors (RMSE) between ground truth and predicted head motions with different models. As expected from CCA analysis and chronograms, the largest

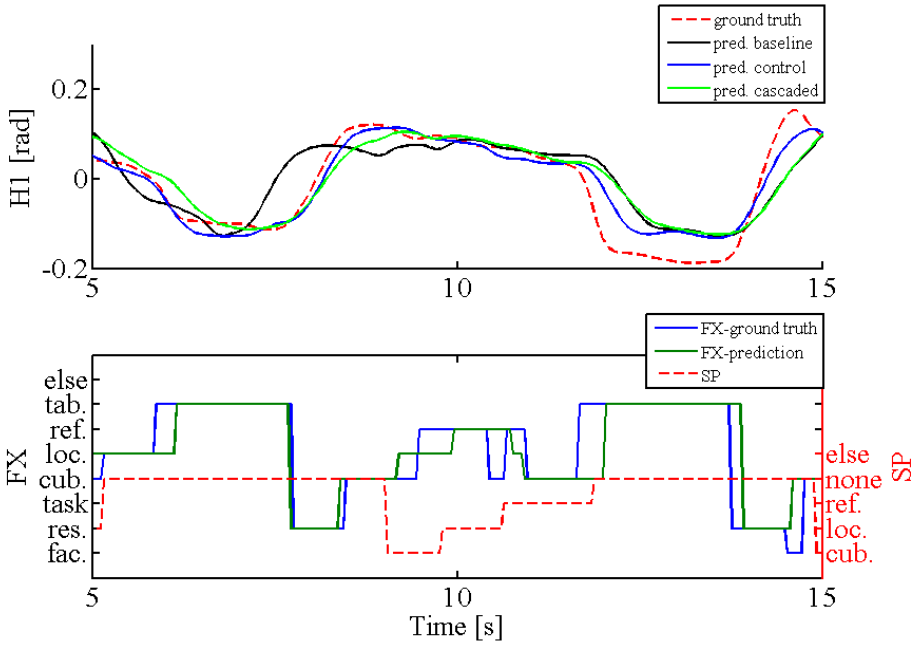


Figure 3.29 – (a) H1 real vs. prediction streams between different models; (b) input streams (FX ground truth& SP) and FX prediction from LSTM1.

Table 3.6 – Root mean square errors (Pearson correlations) between ground truth and predicted head motions with different models.

Models	*H1 [rad]*	*H2[rad]*	*H3[rad]*
Baseline	0.059 (0.84)	0.035 (0.67)	0.066 (0.57)
Control	0.045 (0.91)	0.033 (0.72)	0.066 (0.65)
Cascaded	0.057 (0.84)	0.035 (0.67)	0.065 (0.64)

and lowest RMSE are performed respectively by the *Baseline* (0.059) and *Control* (0.045) models. The *Cascaded* model exhibits an intermediate performance (0.057). Since the CCA of FX, SP and MP are not significantly different for H2 and H3, their RMSE are not significantly improved. Pearson correlations, also given in Table 3.6, corroborate these observations.

In order to compare the micro-coordination patterns, we computed the so-called *coordination histograms* (CH) proposed by Mihoub et al [Mih+16]. In order to conform to their proposal, continuous streams of head motions (H1, H2, H3) are first converted to discrete events by detecting peaks of local maximum velocity. CH are then built by tabulating the time-delay between each event of one given modality and the nearest events from other modalities.

We further compared ground-truth coordinate histogram for H1 with those produced by the different prediction models. Figure 3.30 displays the histogram computed from ground truth vs. CH predicted by the *Baseline*, *Control* vs. *Cascaded* models. Figure 3.31 displays Chi-squared distances between the ground-truth histogram – considered as the target coordi-

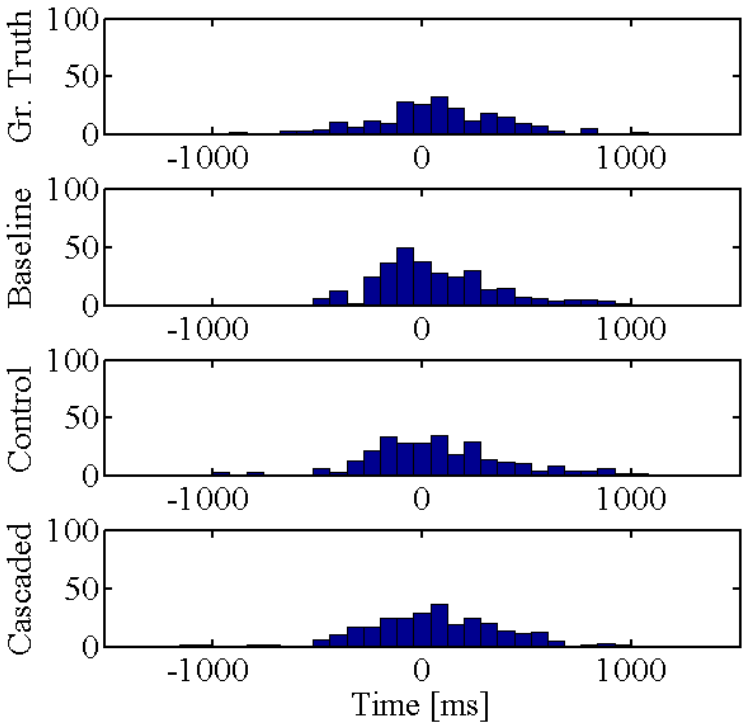


Figure 3.30 – Coordination histograms among H1 and (IU,SP).

nation pattern – with those produced by the different prediction models for the three angles. These figures show that the *Cascaded* model outperforms the *Baseline* model both in terms of accuracy and coordination. This result is partly due to the fact that both *Baseline* and *Control* models are directly driven by triggered events (SP, MP) or FX, while the *Cascaded model* is able to handle the coordination pattern – in particular causal relations – between these events.

3.5.2.3 Discussion

Our experiments evidence that prediction models can benefit from a priori knowledge about causal relations between features. While recurrent neural networks can implicitly construct latent representations using massive data, explicit knowledge given as goals or cost functions help them to build and structure intermediate layered mappings.

Several algorithms to explore the intra- and inter-slice causal relations between observations have been built for DBN and other statistical models. These analysis tools that help to shape probabilistic graphical models (PGM) may be used to automatically structure neural network layers and give ways to shape latent representations with task-related semantic or pragmatic information.

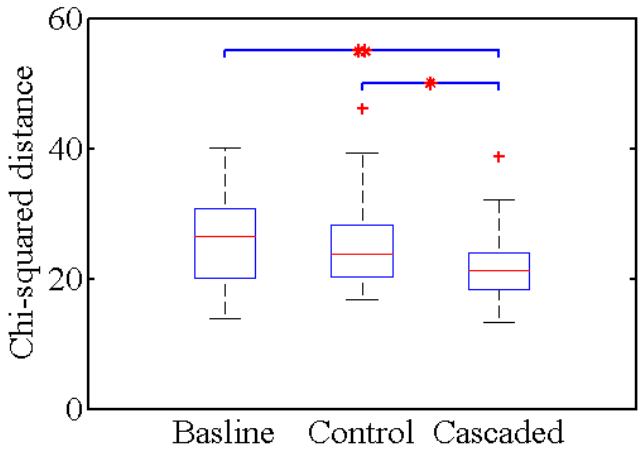


Figure 3.31 – Chi-squared distances of the different prediction models.

We analyze the effect of shifted speech frames using CCA to see the correlation between each features of head movements with the shifted frames. The figure 3.33 illustrates the CCA between H1,H2,H3 and SP, correspondingly. When increasing shifted frame, the CCA of H1 is growing up until 10 frames and get maximum at 0.675. In contrast, the CCA of H2 and H3 reduce when increasing the number of shifted frames.

Although LSTM model can remember events triggered by gaze so that they can drive head motion following the gaze event, it is difficult to deal with subsequent speech events, which are conversely triggered by preceding eye fixations. A way to improve the head motion generation can be done by time-shifting speech, which then becomes the forerunner of head motion. Figure 3.32 displays the H1 RMSE of the baseline model with and without time-shifted input SP frames. SP is here shifted by 10 frames ($\sim$ 0.4 sec): it generates lower RMS compared with the original model. Figure 3.34 shows the H1 RMSE obtained at the optimal epoch corresponding to the different models. Almost all shifted SP generate head motion with lower error. This is well in accordance with the chain of attention driven by gaze – with a rapid eye followed by a slower head motion – that triggers pointin gestures and speech.

Of course, bidirectional LSTM can be used and combined with a soft attention mechanism to optimally probe contextual information (exogenous as well as intentional). But we here consider reactive models that are able to cope with on-line interactive behaviors: the horizon of the contextual information does not extend beyond the current frame.

3.5.3 Conclusions & perspectives

In this section, we propose an efficient solution to structure the intermediate representations built by layered LSTM. We have shown that gaze can be used effectively as a driving signal for head motion generation. This intervention is effective both in terms of accuracy and

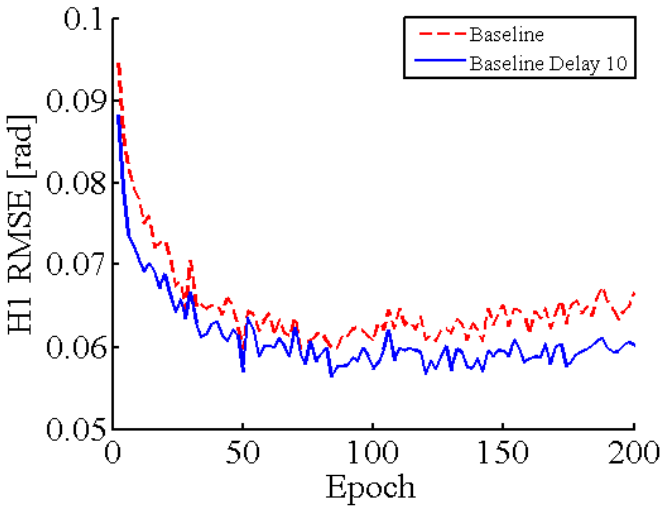


Figure 3.32 – Average H1 RMSE of the *Baseline* model without and with SP shifted frames corresponding to number of training epoch.

coordination patterning.

For now, the cascaded model with HHI data, predicting FX before generating head motion improve just pitch (H1) angle. This could be explained by in the task, there is a larger distance between the tablet (used to info the cube position) and the manipulator space so that the head pitch has a large contribute in gazing. Since the target and manipulator spaces are close to each other and the movement of eyes between these two positions lays in the field of view, the head movement will less contribute to the gaze. Therefore, the H2 and H3 have a smaller correlation with the gaze target comparing to H1. In future, within the immersive teleoperation system, the pilot will perhaps move his/her eyes slower than normal in order to overcome the sensorimotor latency of the oculomotor control. The head movement is thus expected to have a larger contribution to gaze shifts. So, the proposed model could improve all of degrees of head movements by intermediate gaze prediction.

The quality of prediction may be enhanced in several ways. Other contextual information can be used as additional input – precise regions of interest for the gaze, gaze contacts, communicative functions of speech, etc. – as well as intermediate objectives – e.g. eyebrow movements or respiratory patterns. In addition, we did not use the segmentation of the task into IUs because most of these IUs were triggered by gaze or speech events. More complex tasks involving switching between multiple interaction styles with multiple agents may motivate the structuring of the interaction by IUs, notably when alternative cues are used to trigger similar pragmatic frames.

Furthermore, the head motion generation model will be used to drive the head of our iCub-humanoid robot when autonomously instructing human manipulators. We first plan to perform the subjective assessment of our multimodal behavioral model (see [NBE16] for our crowd-sourcing methodology). Another challenge is to adapt this model to multiple ma-

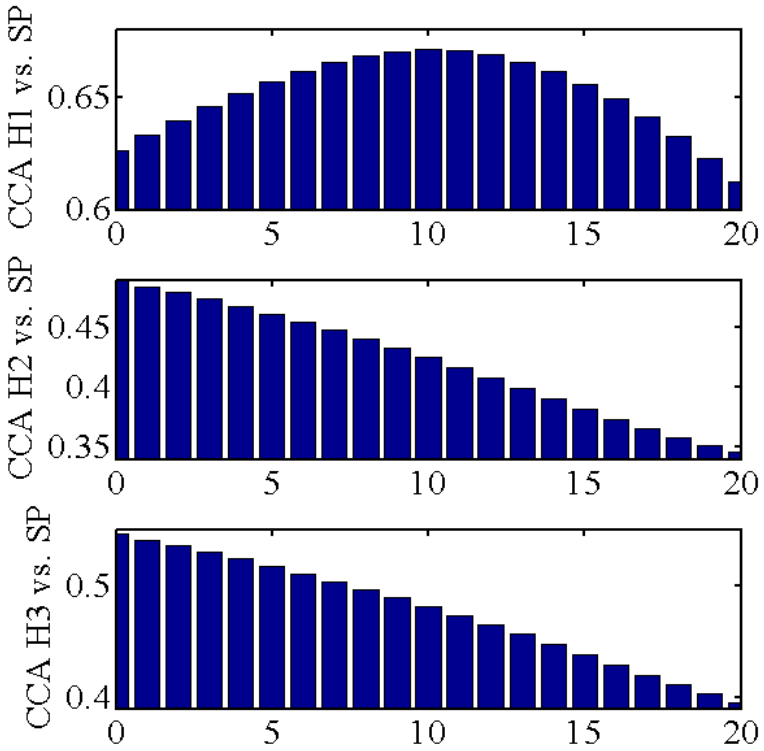


Figure 3.33 – CCA of Hs vs. SP with various number of shifted frame.

nipulators, notably those with motor disabilities. In this case, the behavioral model should both incrementally estimate the best action and the optimal interaction style according to the goodness of fit between the actual and expected behavior of the interlocutor predicted by the joint behavioral model.

3.6 Summary

In this chapter, we present multimodal interactive behavioral models based on recurrent neural networks, namely Long-Short Term Memory (LSTM) RNN for predicting discrete (arm, gaze, backchannel) and continuous (head motion) variables.

The predictions of arm, gaze and interaction units are compared between LSTM for on-line prediction and Bidirectional LSTM (BiLSTM) for off-line prediction with other statistical methods: HMM and DBN. The LSTM behavioral models benefit from extracting contextual information from data, instead of being limited to the boundaries of the hidden states of HMM or the immediate previous frames of the DBN dependency graph. The LSTM methods achieve a better performance than statistical methods with regards to both prediction performance and intermodal coordination.

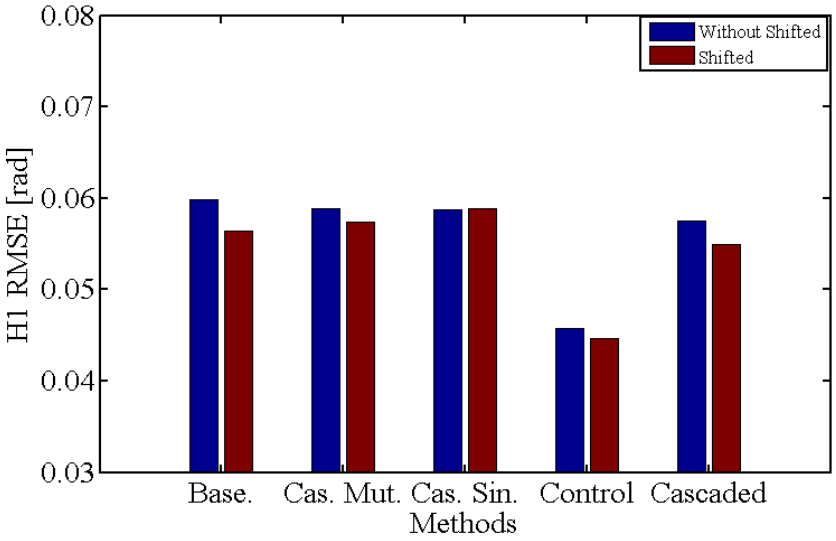


Figure 3.34 – Average H1 RMSE without and with SP shifted frames corresponding to the different models.

For backchannel generation, we compare two methods (CRF vs. LSTM) with and without contextual windows using data from the RL/RI scenario. The LSTM models outperform CRF models. LSTM seems to better capture the relevance and timing of BC in the dialog scenario, where the interviewer usually uses BC to encourage the elderly people to be confident in answering questions.

We also investigated how to use LSTM models to generate continuous head motions. Because of the ability of capturing long-term dependence between latent variables, a LSTM with single layer was used as a base-line model. We used CCA to analyze the PTT interactive data and found that head motions the highest correlation with gaze (FX) and interaction unit (IU) comparing with other features (speech, manipulator's arm, F0). A control model with an additional FX input has a significantly improving head motion generation quality. In order to improve the quality of prediction but kept the same inputs as the baseline model, we built a cascaded LSTM model which uses an other LSTM to predict FX as an input of another that generate head motion. We found that the cascaded LSTM model (pre-trained and fine-tuned parameter) not only improves the head motion accuracy comparing with the baseline, but also has the best coordination.

Chapter 4

Gesture Controllers: Design and evaluation

In chapter 2, we described how to collect human-human interactive (HHI) data and extract useful features for training multimodal interactive behavioral models. Then, we built interactive models (as described in chapter 3) that can generate actions from perception streams in two interactive tasks (*Put That There* and *Selective Reminding Test*).

Note that the interactive models can generate multimodal robot behaviors at both levels: *abstract-level* vs.*skill-level*. The abstract level represents elementary behavioral skills of the target task (e.g."look at (ROI)", "say (text)", "hand-point to (ROI)"), which are described by discrete events. In contrast, the skill-level behavior is related to specific motions such as head trajectories. The feature-level behaviors are generated so that the score they compute can directly command the robot's motor micro-controllers while the skill-level behaviors should trigger specific *gesture controllers* that further convert events into skill-level trajectories. Gesture controllers are thus here the analog of the *gesticon*, the central gesture repository introduced by Krenn and Pirker [KP04], that stored gesture snippets and facial expressions relevant for the generation of dialogue accompanying non verbal behavior of virtual agents.

In this chapter, we focus on designing gesture controllers that can be used to execute the discrete events for our humanoid robot. Building gesture controllers is a fundamental step of developing robot behaviors, which enable us to realize how our robot interacts physically with humans (see Figure 4.1).

For future works, the gesture controllers will be used to build up semi-autonomous as well as autonomous robots to perform the interactive tasks (see chapter 5). Hence, we need to ensure that the events and their synchronization are still perceived correctly by human observers for which they are created. In this chapter, we propose an evaluation framework to spot the robot's faulty behaviors so that they can be redesigned or better adapted. Observing HHI is a good way to design and evaluate the gesture controllers and their relative synchronization. For evaluation, reusing the scores of the HHI data allow to evaluate how the robotic gestures are perceived by human targets, without evaluating the interactive behavioral model at the same time. This leads to focus on corrections of *how* gesture controllers encode elementary skills and *when* events are triggered, and thus partly disentangle execution from planning problems.

In fact, robots have difficulties in performing many actions that tutoring humans can easily perform (e.g. using one's hand to open/close a notebook or use a pen to write). Therefore,

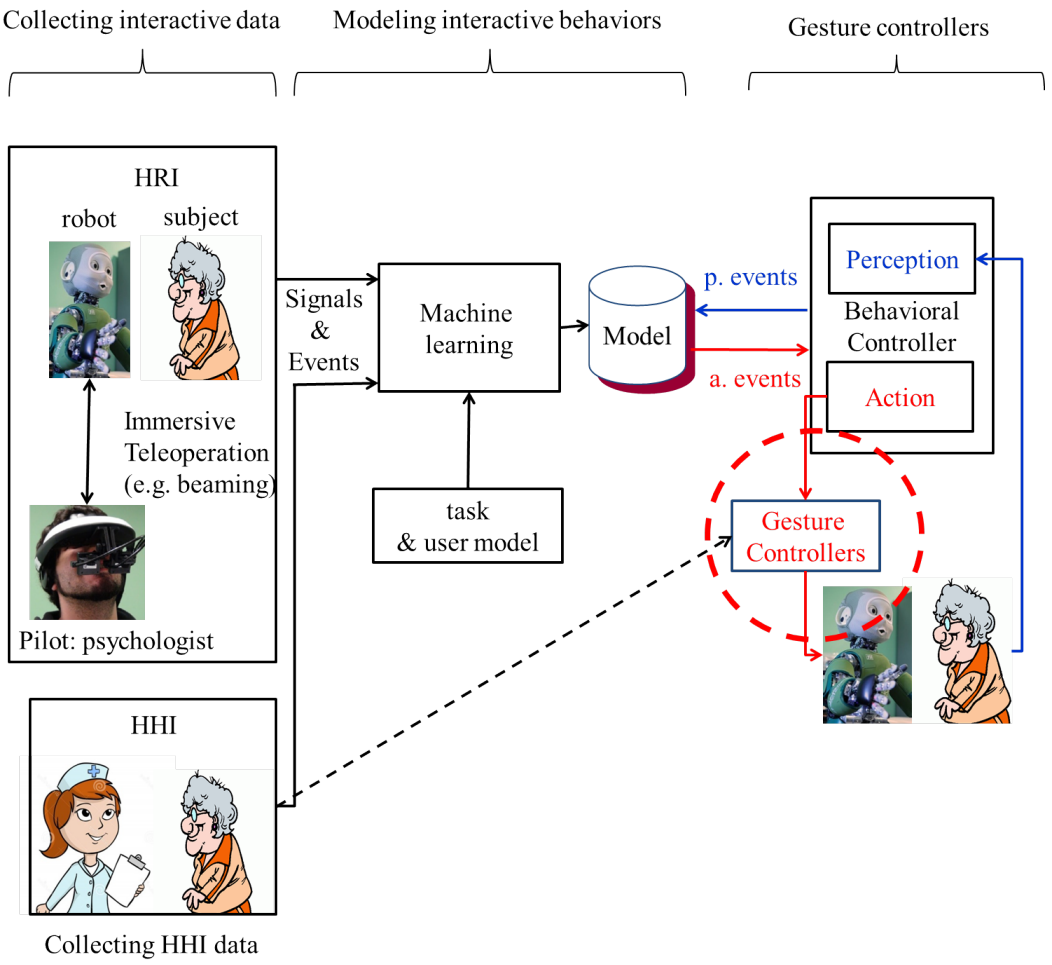


Figure 4.1 – Gesture controllers: design and evaluation. HHI data are used not only to design *gesture controllers*, but also to evaluate the capability of the robot in reproducing coordinated verbal and co-verbal behaviors.

in order for our iCub humanoid robot to perform acceptably the interactive scenarios, some actions of the robot will be changed to better match robot's abilities. In particular, in the *RL/RI* task, instead of opening/closing a notebook to show/hide items, the robot simulates item display and scoring events just by *clicking* on a faked tablet. This chapter covers how we adapt the HHI events to the HRI situation so that the events could be executed easier by the robot while maintaining equivalent semantics of the demonstrated HHI events.

We focus here on the *RL/RI* scenario, which requires the robot to perform much more complex multimodal behaviors and to exhibit more varied social skills than the *Put That There* scenario. We will detail how to adapt the HHI protocol to HRI, design and evaluate gesture controllers for this scenario.

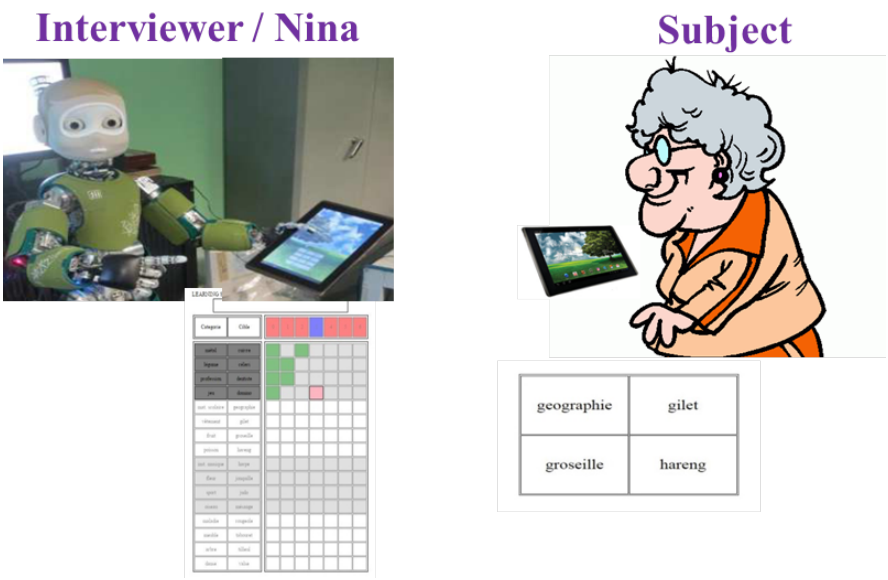


Figure 4.2 – Adapted RL/RI scenario for human-robot interaction: the robot uses a tablet to convince the subject that it drives the display of items and that it effectively takes notes. Another tablet facing the subject displays/hides items according to the robot's needs.

4.1 Adapt the *RL/RI* scenario from HHI to HRI

In chapter 2, we presented the HHI multimodal data, which consists in time-stamped speech, arm/hand gestures and gaze events labeled with their discrete values (e.g. looking at subject's face, tablet . . . ; uttering a text/backchannel with different attitudes; . . .), organized in HHI multimodal scores. Now, we concentrate on developing modality-specific gesture controllers to map these events to robotic actions that a human observer could perceive and understand.

4.1.1 Substituting sheets of paper with displays

Because it is difficult to shape gesture controllers to mimic the interviewer's arm behaviors such as writing on sheets of paper or opening/closing a booklet to show/hide items, the RL/RI scenario is adapted to ease an implementation on the robot. Particularly, instead of using the sheet of papers and the book containing items, an adapted scenario utilizing two tablets is proposed as illustrated in Figure 4.2. A tablet is hold by the robot's left arm and replace the scoring sheets and the other is placed in front of subjects to show/hide the items. In fact, subjects project human skills and capabilities onto agents – including mnesic capabilities – and probably expect the artificial interviewer to still take notes despite its superior memory. Such a behavior is imposed by social rules. Another advantage of the robot's tablet is that it can be used as an augmented display for the operator of an immersive teleoperation system (see chapter 5)

4.1.2 Dealing with response times

Several adapted events from HHI to HRI scenarios are shown in Table 4.1. Actually, when we adapt from HHI to the new HRI scenario, timestamps of adapted events should be changed because of differences in response times and bandwidths between human's and robot's actions. For example, when a robot action takes a longer duration (due to robot's physical limitations) and could not be anticipated, the next action of the robot will be delayed (see Figure 4.3).

In order to minimize this unnatural behaviors, we select robot actions so that their durations will be smaller or equal to those of human actions. As an example, the writing events for taking notes in HHI will be replaced by rapid clicking events in HRI. When performing *preparing score* event, the human interviewer can move his/her arm smoothly and fast from a *rest* position to the score-sheet paper position for writing, while the robot can produce vibrations when moving his arm as the human. In this case, the *preparing score* event will be converted to a *prepare clicking* event where robot's right hand is close to its scoring tablet – hold by its left arm – so that the duration of *preparing click* actions can be performed fast enough.

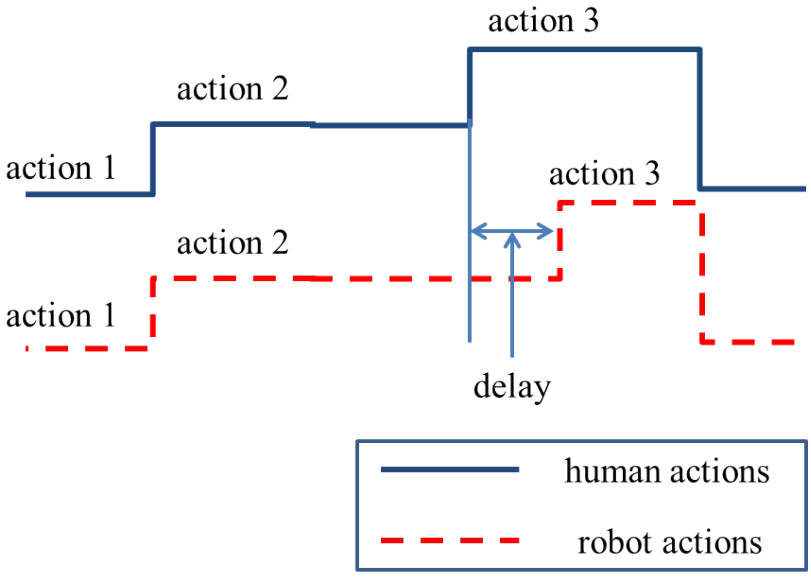


Figure 4.3 – An example of different durations between a robot action and a human action (action 2) when performing the same event. If the duration of robot action is longer than the human one, this can delay the next robot action (action 3).

Table 4.1 – Adapting events from HHI to HRI

Modalities	*HHI*	*HRI*
Gaze target	book containing items sheet of papers	subject tablet interview tablet
Arm gesture	preparing scoring scoring (writing) showing/hiding items	prepare clicking clicking clicking

4.2 Designing gesture controllers

4.2.1 Speech

In the *RL/RI* scenario, the robot will play the role of the interviewer who interacts with Alzheimer patients. This task requires the robot to ask questions and give appropriately feedbacks to the subject about the on-going RL/RI scenario. The robot should also encourage the subjects by giving them rewards and incentives. HHI data show that some sentence constructions are recurrent, and that specific prosodic patterns are recruited [Bai+16]. So we will use the default audiovisual text-to-speech (TTS) system, but with a specifically trained prosodic model.

The built-in TTS system (named COMPOST) was first designed by Alissali et al [AB93] at GIPSA-lab. It controls the several processing stages (text preprocessing, morphological analyzer, part-of-speech tagger, letter-to-sound pronunciation, prosody generation and corpus-based synthesis). The visual component of corpus-based synthesis of French was adapted to the degrees-of-freedom of Nina (vertical movements of the jaw, lower and upper lips, and horizontal movements of the two lip corners; see [Par+15]).

As mentioned in chapter2, speech uttered by the interviewer and the subjects during HHI was transcribed and aligned its phonetic content. The transcription of the interviewer's speech was further augmented with breathing noises – when clearly audible – and discourse markers related to attitudes effectively used in the interaction (assertion, full question, incentive continuation, standard continuation, unmarked utterance).

From the recorded speech data, a specific model for rhythm and melody was trained using the SFC [BH05], a trainable prosodic model developed in the lab. This model considers the prosody as the superposition of multi-parametric contours as shown Figure 4.4. Each contour encodes a specific function: attitudes at the level of the utterance (green), syntactic dependency relations between syntactic constituents at the level of phrases (blue) and emphasis at the level of words (orange). The shape of these contours are learned via an analysis-by-synthesis process that trains feed-forward neural networks to associate positional features of each syllable with its F0 pattern and lengthening factor (see recent developments performed by Gerazov et al [GB18]; [GBX18]).

An excerpt of the interviewer's speech – ready as robot's speech input – during her inter-

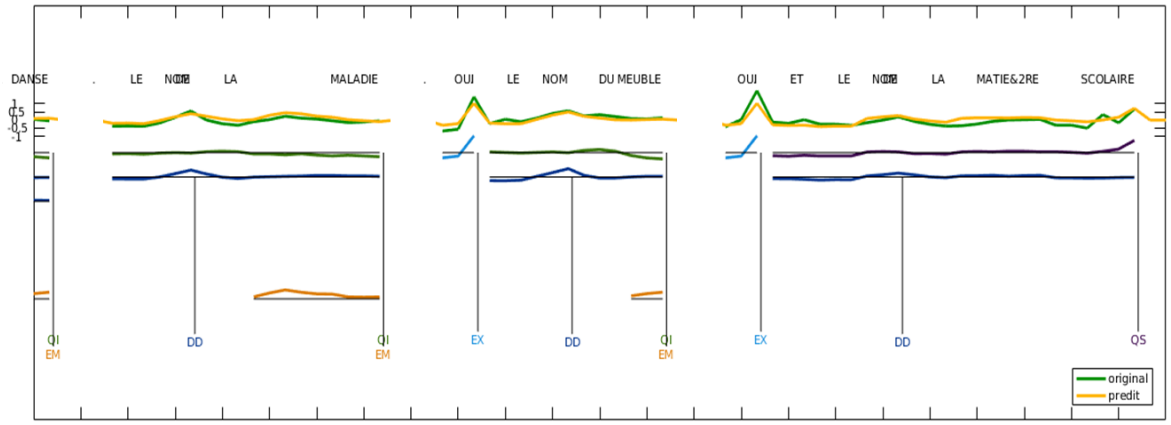


Figure 4.4 – Predicting prosody with superposition of functional contours: top for melody and bottom for rhythm. There are several specific functions in different levels: utterance in green, syntactic of phrase in blue and words in orange.

action with the first subject is given below:

```
[...]
<debit;tm=3.8> #CT <hr;duree=266> hh ALORS NOUS.
<debit;tm=4.6> #CTp ON VA APPRENDRE SEIZE MOTS ENSEMBLE.
<debit;tm=6.9> #CT <hr;duree=261> hh ON VA LES APPRENDRE PETIT A&2 PETIT, QUATRE PAR QUATRE.
<debit;tm=9.7> #DC DONC, IL FAUT BIEN VOUS CONCENTRER POUR BIEN RETENIR LES MOTS QUE JE VOUS MONTRE.
<debit;tm=13.4> #QS D'ACCORD?
<debit;tm=14.2> #DC <hr;duree=434> hh ALORS VOICI LES QUATRE PREMIERS MOTS.
<debit;tm=16.8> #QI <hr;duree=309> hh EST-CE QUE VOUS POUVEZ ME LIRE LE NOM DU POISSON?
<debit;tm=21.2> #EX OUI.
<debit;tm=21.8> #QI LE NOM DU VE&3TEMENT?
<debit;tm=24.7> #EX OUI!
<debit;tm=25.2> #QI LE NOM DU JEU?
<debit;tm=27.7> #QS ET LE NOM DE LA FLEUR?
<debit;tm=30.0> #QS LA FLEUR, REDITES-MOI?
<debit;tm=32.0> #DC D'ACCORD.
<debit;tm=33.0> #EX ALORS, J' ENLE&2VE LES MOTS.
<debit;tm=34.4> #QI EST-CE QUE VOUS POUVEZ ME REDIRE LE NOM DU POISSON?
<debit;tm=37.6> #QI <hr;duree=597> hh LE NOM DU VE&3TEMENT.
<debit;tm=40.4> #QI <hr;duree=535> hh LE NOM DU JEU.
<debit;tm=46.1> #QS <hr;duree=381> hh ET LE NOM DE LA FLEUR.
<debit;tm=49.2> #DC OUI .
<debit;tm=49.9> #QI ET CE JEU ALORS?
<debit;tm=52.7> #QS IL EST PARTI?
<debit;tm=54.8> #DC ALORS, JE VOUS REMONTRE.
[...]
```

In the excerpt above, the attribute `tm` of the global variable `debit` sets the triggering time of the *say* command of the speech controller. Objects starting with `#` set the prosodic attitude of the following sentence(s) – `DC` for assertion, `QS` for full question, `EX` incentive continuation, `CT` standard continuation ...– They overwrite the default modal prosody given by the final punctuations. The attribute `duree` of the phone `hr` sets the duration of a breathing noise. The phone `hh` is a glottal stop with a default duration of 50ms. Orthographic input is provided via uppercase letters, accents being provided by numbers preceeded by `&` – `&2` is grave, `&6` is

acute, &3 is circumflex ...

4.2.2 Arm gestures

While the human interviewer was scoring and displaying word items using sheets of paper, we decided to use tablets to take notes and display items while pretending to tick boxes (see Figure 4.5). Note that subjects are easely fooled by such a *magic finger*: Hood et al [HLD15] similarly convinced children that the Nao robot was able to perform cursive hand-writing! This sensorimotor integration is all the more credible if the tracing/gesture synchronization is neat.

(a) scoring.

(b) resting.

Figure 4.5 – Robot's arm.

Arm displacements and finger clicks are triggered to synchronously trigger display on the subject's tablet (show/hide items) and take notes (monitor responses, whether correct or not). The arm gesture controller uses the iCub Cartesian Interface [Pat+10], which enables the control of the robot's arm by providing the desired position and orientation of one end-effector (here the index finger of the right hand) directly in the 3D space. Our arm controller also provides task-specific movements: preparing to click, clicking, and going back to rest position. Figure 4.5 illustrates the position of robot's right arm while scoring and resting. In the experiment, the left arm remains fixed. It holds the scoring tablet, while the right arm movements are adapted so as to follow – as closely as possible – the timing of the original writing actions of the human interviewer.

4.2.3 Gaze

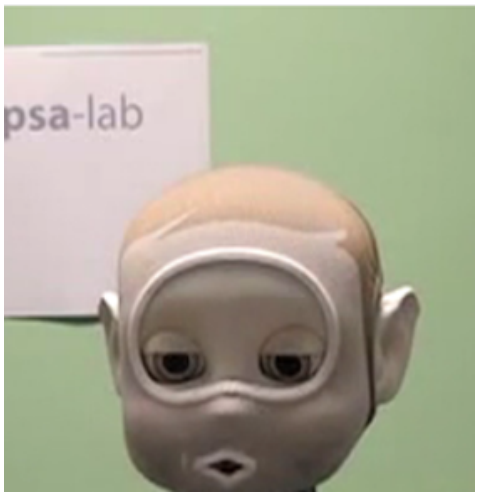
We distinguish three main regions of interest for the interviewer's gaze: (1) the subject's face; (2) the scoring tablet (i.e. the scoring sheet and chronometer used in the original HHI); (3) the subject's tablet (i.e. the notebook used in HHI demonstrations). Note that all arm gestures are performed with visuomotor supervision: since robot motion is often slower than human motion, all arm motions are preceded by one fixation towards the target, if any, and accompanied by gaze smooth pursuit till completion. This visuomotor supervision supersedes the original fixation patterns.

The gaze gesture controller uses the iCub gaze controller [Ron+16], which provides direct control of saccades, fixations and smooth pursuit while implementing the binoccular vergence, the oculo-collic and vestibule-ocular reflexes. These gestures can be performed by a parametrized combination of neck and eyes movements. For simplicity, the Cartesian gaze controller is provided with the 3D position of the current region of interest and a fixed contribution of neck movements of 50%. Figure 4.6 presents the final robot's head position for two targets: (a) looking at subject's face, (b) looking at scoring tablet.

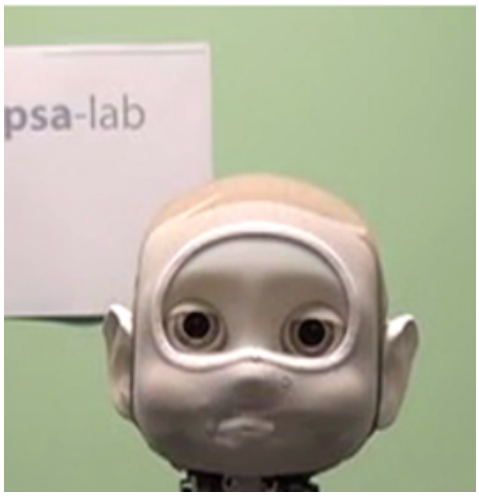
(a) looking at subject's face.

(b) looking at scoring tablet.

Figure 4.6 – Two examples of robot's gaze.



(a) gaze looking down.



(b) gaze looking straight.

Figure 4.7 – Robot's eyelids.

4.2.4 Eyelids

Although we did not track eyelids' movements, we developed a specific eyelids gesture controller in order to provide Nina's behavior with additional socio-communicative cues such as

blinking as well as redundant cues such as the coupling of eyelids aperture with eyes elevation [FBE15]; [Bai+06] as well with speech articulation [Bad+02]. Figure 4.7 illustrates the coupling of eyelids aperture with eyes elevation. Note that the lack of such coupling of eyelids aperture with eyes elevation may result in wrong interpretation of the associated facial expression [San+07].

4.3 Evaluating gesture controllers

Can we ensure that these complex and coordinated behaviors are correctly perceived and interpreted by human subjects? Evaluating a fully autonomous robotic system would cumulate both weaknesses of the gesture controllers and the behavioral controller – i.e. the event generator – at the same time. We would like to evaluate the ability of the gesture controller to reproduce the interactive behavior of interviewer independently of the high level behavioral controller. Therefore, the gesture controller will be first driven by events extracted from the human-annotated events from HHI data , further adapted for HRI as described above.

Since subjects can not both experience and rate the interaction on-line, we thus asked third parties to rate the rendering of a multimodal HHI score "replayed" by our robotic system in order to check if the reconstructed robot's behavior is still relevant and if the mapping between discrete events and gestures are correctly performed by our gestural controllers.

4.3.1 Evaluation of HRI systems: state of the art

Quality of social HRI is often assessed from three perspectives [You+11]: (1) visceral factors of interaction, (2) social mechanics and (3) social structures. The first perspective focuses on *instinctual aspects* such as fear, excitement, joy, happiness. Uncanny valley is one of highlighted example of this perspective in which the shape, speed, and patterns of a robot's movements contribute to visceral reactions [Mor70]. The second perspective concentrates on *social techniques* used in the interaction, such as range of gestures: facial expressions and body language, eye-contact rules. The final perspective focuses on *social relationships* over long period of time. The two first perspectives serve as guidelines for our evaluation paradigm: we used *post-hoc* questionnaires to question visceral reactions and *on-line* detection of violations of social techniques.

Most subjective evaluations of HRI behavior have been performed using questionnaires, where subjects or third parties are asked to score specific dimensions of the experienced interaction on a Likert scale, after having watching it. Fasola et al [FM13] rated several aspects such as pleasure, interest, satisfaction, entertainment and excitation. Huang et al [HM14] assessed a narration humanoid robot along several dimensions such as immediacy, naturalness, effectiveness, likability and credibility. Zheng et al [ZWM15] compared control strategies for robot arm gestures along dimensions such as intelligibility, likeability, anthropomorphism and safety. Scholtz and Bahrami [SB03] focused on predictability of behavior, capability aware-

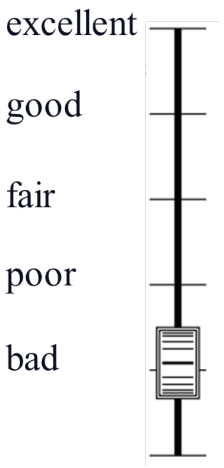


Figure 4.8 – Schematic model of continuous quality rating with scale [HK99].

ness, as well as interaction awareness and global user satisfaction. Nomura et al [Nom+05] analyzed the feeling of visitors about robots exhibitions (greeting and guiding the visitors and so on) in several aspects: interest, friendliness, effectiveness, anxiety toward interaction and anxiety toward social influence. Heerink et al [Hee+09]; [Hee+10] published a rather extensive questionaire for measuring acceptance of an assistive social robot, in particular by elderly users.

Although delivering very useful information, notably for sorting between competing control policies or settings, these questionnaire-based evaluations provide developers with poor information about how to correct faulty behaviors since the evaluation is performed off-line and questions address global properties of the entire interaction.

4.3.2 Designing and performing on-line vs off-line evaluation

On-line evaluation methods have been proposed for audio [HK99] and video [HR95]. In these works, raters continuously indicate the perceived strength of sound or image quality by moving a slider along a graphical scale (shown in Figure 4.8). Similarly, Tanaka et al [Tan+06] proposed a continuous evaluation method in order to evaluate long-term interaction relationship: they ask participants to rate continuously in 1-5 scale when watching a recorded video in which the QRIO robot is dancing with children in canned and interactive conditions. The evaluation results showed that children loss interest on the robot as time progressed.

Following the procedure proposed by de Kok & Heylen [KH11], we opted for a method that enable raters to signal faulty events, since the HRI behaviors are essentially controlled by events. We thus designed an on-line evaluation technique that consists in asking raters to immediately signal faulty behaviors by pressing on the “ENTER” bar of their computer when they just experience them. Following Kok & Heylen, we will use the term *yuck responses* to name these calls for rejection.

Figure 4.9 – The *Nina* robot from the subject's perspective.

Since raters cannot both experience and assess an interaction, we ask them to put themselves in the place of the subject who has originally experienced the interaction (i.e. to feel the experience without actually performing it). In order to gather a significant amount of yuck responses for a set of identical stimuli, we here ask our raters to evaluate the replay by the robot of the multimodal behavior, originally performed by our psychologist in front of one unique subject. We in fact filmed the robot's performance from face – giving the impression of a first-person view – while fed by the multimodal score of the original situated interaction (arm gestures, head movement, gaze...). For now, the only original play-backed behavior is the subject's speech input. The camera remains fixed at the mean position of the location of the eyes of the subject: this subject's perspective is shown in Figure 4.9). The raters can see the robot facing them, but not the patient that they "replace". They can hear the robot, as well as what the subject says: they are spectators, but occupy the subject's seat.

A general framework of designing and evaluating gesture controllers is shown in Figure 4.10. The purpose of the framework is to locate faulty behaviors and correct these behaviors before a new evaluation. The framework is like an ecosystem that enables us to analyze and enhance the faulty behaviors of gesture controllers as well as suggest missing events in the HRI score.

In each evaluation, peach articipant also filled a post-hoc questionnaire just after the experiment that ask judgments (five-level Likert scale) on nine points:

1. Did the robot adapt to the subject?
2. Did the subject adapt to the robot?
3. Did you feel relaxed?
4. Did you feel secure?
5. Was the rhythm of the robot's behavior well adapted?

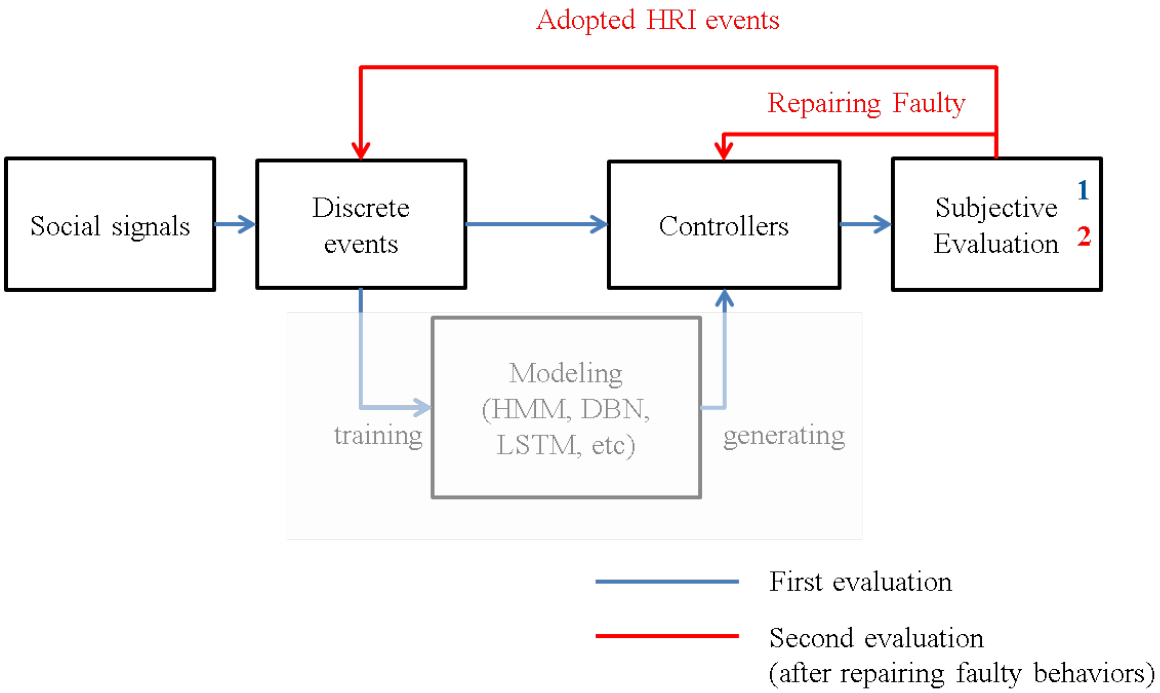


Figure 4.10 – General framework of evaluation gesture controllers. Blue lines illustrate processes of the first evaluation. Red lines illustrate the second evaluation after repairing detected faulty behaviors of the first evaluation.

6. Was the interaction pleasant?
7. Was the multimodal behavior appropriate?
8. Did the robot pay attention while speaking?
9. Did the robot pay attention while listening?

4.3.2.1 The first evaluation

For our first experimental assessment [NBE16], we created a website [1] where we ask people to look at a first-person video and to press the "ENTER" key anytime they feel the robot behavior is incorrect. This on-line evaluation task is preceded by a quick screening of subjects (age, sex and mother tongue) and a familiarization exercise. 50 French natives (26 males / 24 females) performed the first evaluation. The age of the participants is 32 ± 12 years.

The cumulated yuck responses provide a time-varying normalized histogram of incorrect behaviors (cf Figure 4.11). The maxima of the density function are cueing time-intervals for which a majority of raters estimate the behavior is inappropriate or hinders the interaction. Further diagnostic of what cues cause these faulty behaviors are later performed by roboticists and system designers, seeking for incongruous behaviours around $200ms$ before the peaks, i.e;

[1] http://www.gipsa-lab.fr/~duccanh.nguyen/assessment/

taking response time into account.

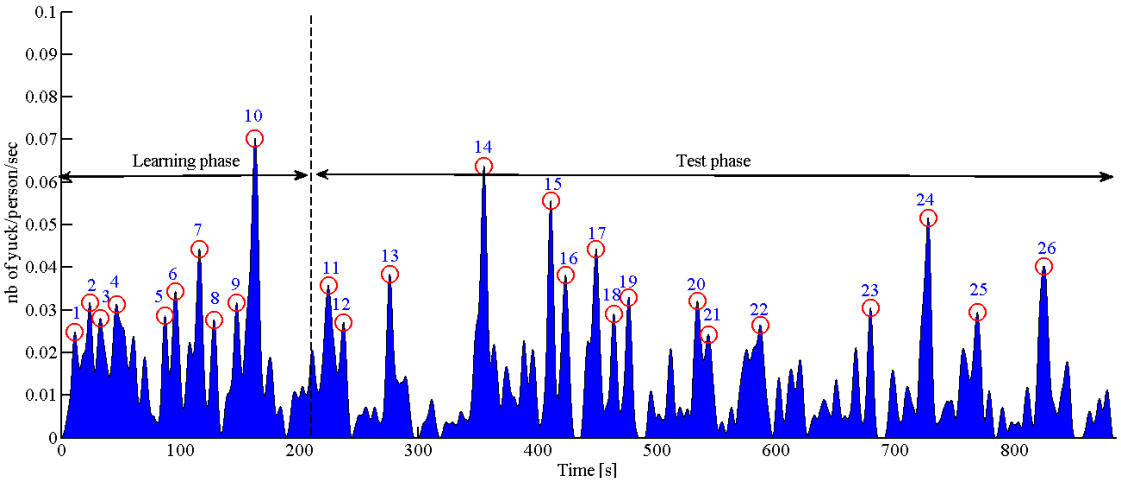


Figure 4.11 – The yucking probability as a function of time for first by participants.

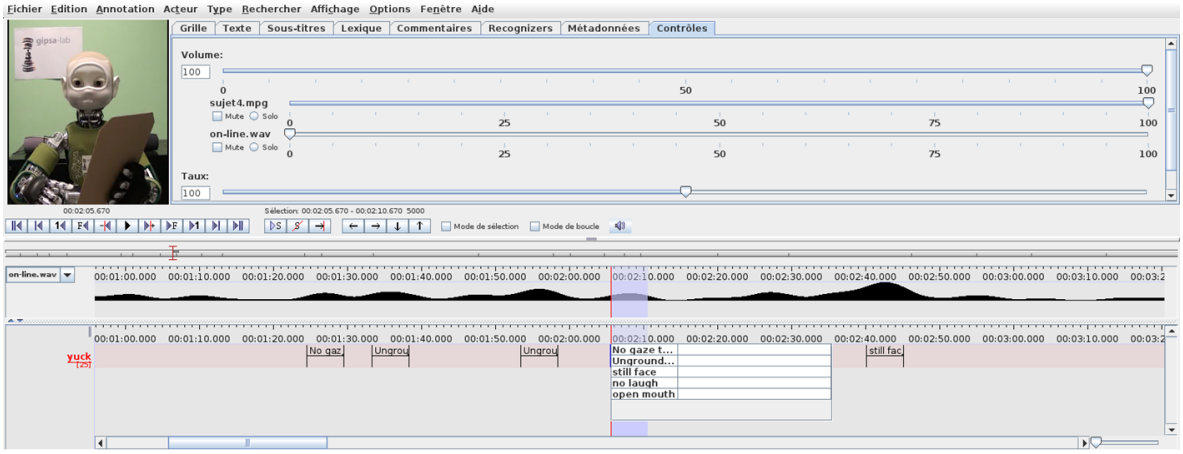


Figure 4.12 – Density of yuck responses for our replayed interaction. Each yuck response is weighted by a Hanning window of 5s in order to smooth the density of responses and overlapped/added to the others. Maxima of this time-dependent histogram reveal multimodal behaviours that are judged inadequate by a majority of raters.

Most of the time, the maxima of this density function have clear interpretation. We then use Elan (see Figure 4.12) to pair these majority yuck responses with multimodal events. Here are the 26 most signaled events:

unknown click 1, 4, 6 & 7: the robot performed clicking gestures on its tablet to show or hide items onto the subject's display that was not available to raters (!). Such ungrounded gestures are thus perceived as distractors by subjects. These yuck responses are located at the beginning of the interview, during the learning phase.

missing gaze 2, 3, 5, 9, 17, 20, 21, 22, 25 & 26: participants also detected that gaze towards the subject was missing or too much delayed with reference to the interviewee's answers to questions or when delivering instructions. While such a behavior is quite legible when

performed by the interviewer – who did not want to interfere with the subject's thoughts – but seems completely unacceptable when performed by a SAR, whose intentions are much less readable.

The lack of gazing at subject face is also caused by delayed clicks. In fact, in the adapted scenario, the robot performs *clicking* actions on its tablet instead of writing on a notebook. As a natural behavior, we opted for a simple rule: when the robot's arm clicks to the tablet, the gaze should be directed to the tablet. However, these clicking actions are often delayed because of wait-motion-done options which makes *preparing clicking* actions longer than expected (see Figure 4.3).

remaining still 10, 11, 12, 13, 16, 18, 19 & 24: the robot remains still – with the exception of quasi-periodic blinks – for too long, notably during periods of poor interactive activity of the interviewee such as reverse counting or covert thinking. This absence of input observations results in no generated movements.

open mouth 14 & 15: these particular misbehaviors are explained by the persistence of a large mouth opening well after finishing speaking. This failure is now identified and has been corrected: it was due to a faulty audiovisual segment that was improperly articulated during a silent pause.

no empathy 23: In several place, the subject joked and laughed. The lack of SAR response to this strong call for social support during episodes of embarrassment is rightly penalized by raters.

unknown 8: We did not found any obvious explanation for this particular yuck response.

4.3.2.2 The second evaluation

Following the first evaluation, we tried to remedy to these problems, in particular by adding a default scanning gaze pattern towards the interviewee's face, correcting visual turn taking/closing gestures and implementing a better synchronization between the subject's responses and scoring gestures performed by the robot.

Table 4.2 – Causes of yuck behaviors of the first evaluation and modifications for the second evaluation

First evaluation		**Second evaluation**
Yucks	*Caused by*	*Modification*
missing gaze - clicking delay	wait-motion-done (WMD)	disable WMD
remaining still	poor interactive activity	gaze at previous state
open mouth	articulatory mapping error	correct mapping data
unknown click	show/hide items	no modification
no smile	no robot smile yet	no modification

Table 4.2 lists the main detected faulty behaviors of the first evaluation such as articulation errors (open-static mouth), still-remain behaviors, ungrounded clicking gestures for showing/hiding items, no smile and clicking-delay.

After correcting these faulty behaviors, we performed a new experimental assessment using the same experimental protocol. The second experimental assessment was performed by 46 French native subjects (16 males, 30 females, 36±16 years). 8 of these raters already participated in the first assessment.

4.3.3 Comparing the two evaluations

4.3.3.1 Yuck responses

In the second experiment, we remedied to these faulty behaviors by adding extra-rules to our gesture controllers. For example, in order to avoid immobility due to periods of poor external stimulation, the gaze controller automatically randomly loops on the two last regions of interest when the delay from the last fixation exceeds 3 sec. With this rule, the number of yucks at timestamps 10, 11,12, 13, 16, 18, 19 and 24 are significantly reduced as shown in Figure 4.13.

However, this randomization should not be equally distributed and should favor the subject's face, since the participants still complain about its lack of engagement with the human subject (e.g. around peak 11,12,13 in counting task). This problem will be suppressed by systematically adding the subject's face to the current attention stack and favoring this region of interest in the gaze distribution.

In the first evaluation, yucks such as those occurring at timestamps 2, 3, 5 were due to the wait-motion-done setting. In the re-design, these faulty behaviors have been removed by disabling the wait-motion-done option that discard any new command while the current gesture has not reached its target according a given precision. We supposed that viewers are able to decode intentions and authorize the interruption of the robot's movements . This policy is efficient: the yuck responses at landmarks 2,3,5 are significantly reduced in the second evaluation. The yucks at landmarks 14 and 15 were repaired by forcing the closing gesture at the end of phonation.

Although many of the faulty behaviors are suppressed, several faulty detections still remain while some new yucks emerge from the background, notably the absence of expressiveness, e.g. emphatic responses to subject's embarrassment or head nodding normally associated with incentives, respectively cued by yellow vs. cyan extrema.

We compared the probability distributions of yucking for the first vs. second assessments. We also further distinguished between subjects who performed both assessments. The average yucking frequency is respectively 0.013, 0.007 and 0.007 yucks/s for the three groups. The difference of the average yucking frequency between the first vs. second assessment is statistically significant, whether subjects participated to both experiments or not. This clearly shows that we effectively succeeded in resolving some of the faulty multimodal behaviors since the average yucking probability is divided by a factor of two between the two evaluation sessions.

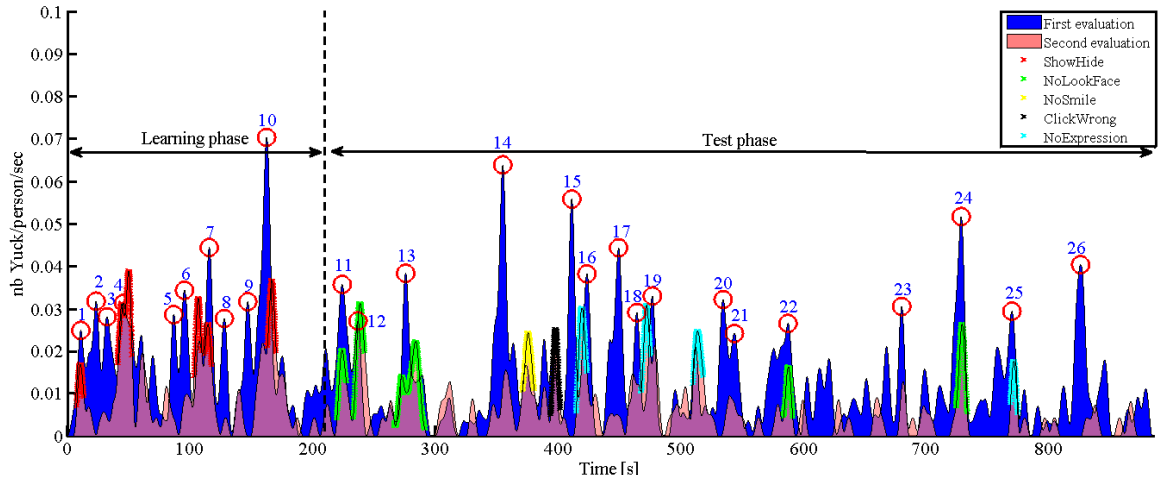


Figure 4.13 – Comparing the yucking probability as a function of time for first vs. second assessment by the subjects.

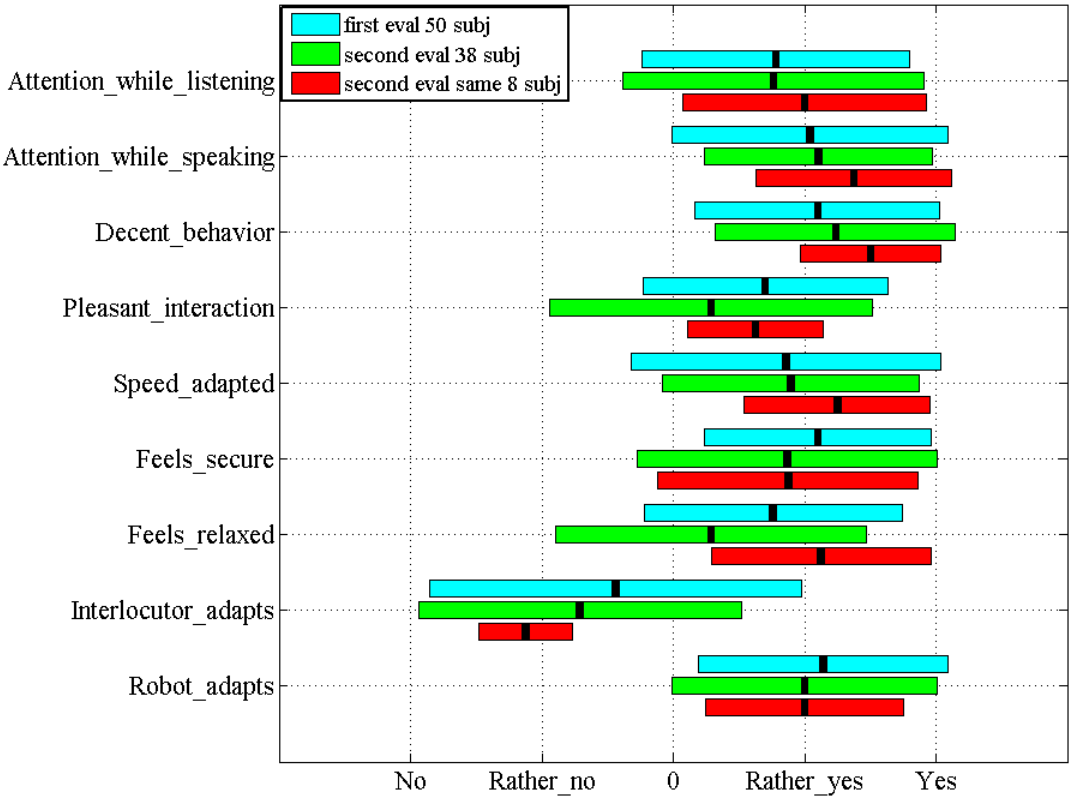
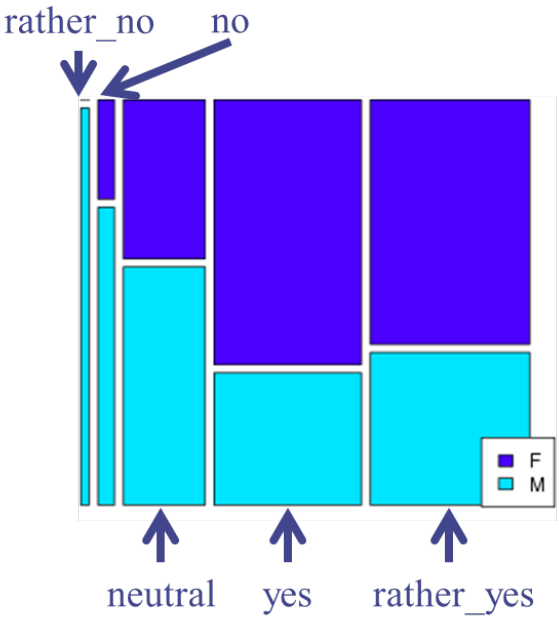
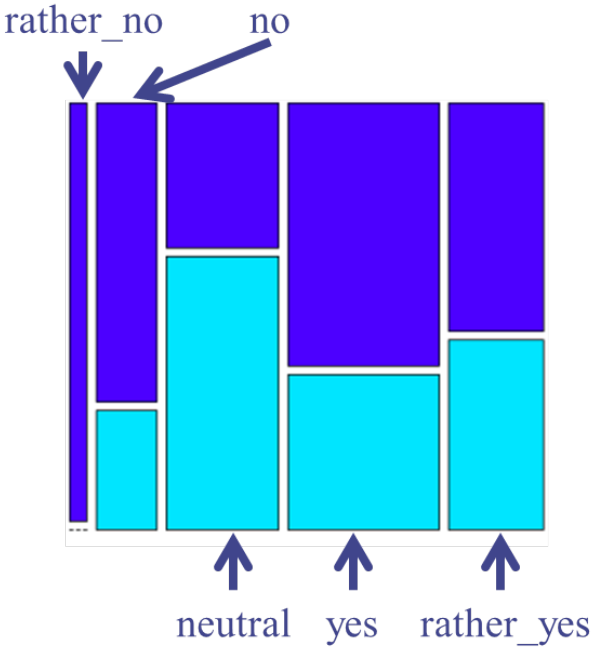


Figure 4.14 – Comparing subjective ratings according to conditions (same conventions as figure 4.13).



(a) Robot adapts



(b) Feels relaxed

Figure 4.15 – Overall repartition of ratings to questions 1 and 3 according to sex

4.3.3.2 Subjective ratings

We also compared subjective ratings from the first vs. second assessment (see figure 4.14). While the new behavioral score results in an effective decrease of the yuck responses – and the rating of *descent behavior* effectively improves – most other off-line subjective ratings degrade. Likelihood ratio tests comparing the combined multinomial model RATINGS $\sim$ SEX+SESSION+EXPOSURE with the individual models RATINGS $\sim$ SEX+SESSION, RATINGS $\sim$ SEX+EXPOSURE and RATINGS $\sim$ EXPOSURE+SESSION show that SEX significantly contributes to the ratings of questions 1 and 3 as shown in Figure 4.15 (females being less convinced by the robot's adaptation capabilities but more relaxed than males). In addition, the second version has significant contributions on two ratings, i.e. *feel relaxed* ($p < 0.02$) and *pleasant interaction* ($p < 0.09$). This means that people feel more relaxed and the robot was rated as more friendly in the first evaluation.

4.3.3.3 Comments

In the free comments, some raters of the first evaluation campaign mentioned the rather directive style of our female interviewer and the absence of emotional vocal and facial displays of our SAR – e.g. laughs and smiles. While most raters of the second evaluation campaign underly the quality of gaze behavior, the majority criticize the poorness of emotional displays: "robot without human warmth!", "why robots never smile?", "[the robot] does not react to humor", "in its behavior, I sometimes felt boredom or weariness", etc. It seems that the increased behavioral quality and appropriateness also increased the participants' expectations. As they have the impression that the robot is reactive, aware of the situation and monitors the interaction task in an appropriate way, they can allocate more attentional resources to the social and emotional aspects of the interactive behavior.

These critical reviews concur with Masahiro Mori statements [Mor70] about the uncanny valley, or perhaps more likely the uncanny cliff hypothesis [Bar+07] that postulates that the likability of robots may evolve on an uncanny cliff without necessarily falling in the valley. The challenge is then to maintain performative, socio-communicative and emotional behaviors at the same level of acceptability.

Moreover, these experiments show the limits of HHI-to-HRI transfer learning: multimodal behaviors exhibited by human tutors may not be fully acceptable by SAR. Here, while the neuropsychologist is quite licensed to concentrate on her score sheet while the subject is performing a counting task or trying to retrieve an item from memory, such a casual behavior is associated with carelessness and coldness from a SAR. This has to be confirmed by asking our web subjects to rate – using the same methodology – the behaviors of a panel of human neuropsychologists. Cormons et al [CDP16] are notably comparing bahaviours of practitioners as a function of their curriculum: nero-psychologists, clinicians vs. speech therapists.

Goetz et al [GK02] found that robots with playful behaviors are usually rated more positive than robot with serious personality, but people followed the instructions of “serious robots"

for longer. Therefore, the elderly people may not interact with the robot if they do not behave seriously. Therefore, we will still keep this version of the gesture controllers for further works.

4.4 Conclusions

In this chapter, we illustrate how to transfer action abilities from HHI to HRI by adapting the *RL/RI* scenario. In fact, the robot has difficulty in reproducing several actions that the human interviewer usually use when performing the task (e.g. writing, opening/closing a book). Therefore, some actions of HHI scenario were modified to be suitable for the robot's abilities. In particular, instead of using sheet of papers to note scores and using the book to show/hide items, the adapted scenario uses two tablets so that the complex actions (e.g. writing, opening/ closing the book) of the human interviewer will be converted to simple actions for the robot by just monitoring the tablet (e.g. clicking, prepare clicking).

Then, gesture controllers (speech, gaze, arm, eye-lids) were designed in order for the robot to be able to execute action events of the adapted scenario. In the following chapter 5), the gesture controllers will be used to drive semi-autonomous as well as autonomous robots.

An evaluation framework was also proposed to detect the robot's faulty behaviors as well as to adapt HHI scores to HRI (e.g. adding more gaze at subject's face in the counting tasks). We perform here two evaluations: the first one for detecting faulty behaviors and the second one to verify the effectiveness of the corrections. The yuck responses of the second evaluation are significantly reduced compared with those of the first one. This demonstrates the effectivness of our evaluation framework. We however found that, after repairing the faulty behaviors, people seem have higher expectation on the robot's behaviors (e.g. they expect the robot to smile; they feel the robot more serious). We will still keep the second version of gesture controllers for future work because we believe that performance and reliability is more important in these short-term task-oriented interactions than pleasure.

Chapter 5

Towards Autonomous Robots and Evaluations

In previous chapters, we described how to collect HHI data of two interactive scenarios, process the data to scale the multimodal score to the robot's sensorimotor abilities and train interactive behavioral models which are able to generate action events from perception streams. We have interactive models to generate gaze, arm events and head motions for the *Put That There* scenario, and backchannels for the *RL/RI* scenario. With these multimodal interactive behavioral models, we are getting closer to build an autonomous robot that can interact with human subjects. This chapter presents our ongoing works to achieve **autonomous robots** that can perform automatically the two short-term interactive scenarios.

An autonomous robot, here, is a robot that can perceive its environment and has the ability to generate actions automatically to interact with human subjects in order to perform the situated tasks. We construct such an autonomous robot by several main modular subsystems:

- **Perception Modules** to enable the robot to percept correctly human environments (e.g. human actions/ behaviors).
- **Interactive Modules** to generate adequate actions from perception streams.
- **Gesture Controllers** to execute actions generated by the Interactive Models
- **A task model** to manage sub-tasks to guarantee that the scenarios are performed completely.

For the quality of the whole autonomous system, each module should be built and evaluated adequately. We develop each module step-by-step in a *"scaffolding"* way so that we can focus on improving and verifying the quality of each module. This chapter covers which modules are available and how to evaluate them in a proper way, as well as which modules are still unavailable and how to build them in the future.

The next section will present works in progress to build an autonomous robot that will collaborate with humans in the *PTT* scenario. For the *RL/RI* scenario with almost available modules, we will present **a first autonomous architecture** based mainly on the results we have done so far. This autonomous robot architecture, designed for a **basic autonomous** robot, will be used as a baseline to improve its behaviors in the future. Next, we analyze

the remaining challenges of building the higher-level cognitive abilities of the social robot and propose approaches for solving step-by-step these challenges.

5.1 Towards an autonomous robot performing the *Put That There* scenario

This section focuses on designing an autonomous robot that can perform the *PTT* in the future. We describe here modules that have been achieved so far and specifications for the missing modules. Then we present our strategy to evaluate available modules and discuss our perspectives on building remaining modules.

5.1.1 Required Modules

Four main modules to run an autonomous robot performing the *PTT* are detailed as following (shown in figure 5.1 (c)):

- **Interactive models** were built to generate gaze, arm events and continuous head movements (see chapter 3).
- **Gesture controllers**

  Gaze and speech controllers are designed almost in a similar way to gesture controllers replicating the *RL/RI* scenario (see chapter 4).

  The arm controller is designed to enable the robot moving from a rest position to pointing at a target position (e.g. reservoir, target location or reference cube). To make a precise pointing gesture that matches with gazing at the same position, we perform a sensorimotor calibration in order to map precisely from gaze fixation to arm pointing target. The calibration process is described in detail in appendix A.
- **Perceptions Modules**

  To run an autonomous robot performing the *PTT* scenario, complex perception modules are required, however, their building is outside the boundaries of this thesis. Here, we give some discussions about the requirements of these modules and our perspective to achieve this.

  In order to capture **inter-coordinations** between the robot and a manipulator, inputs of the interactive behavioral models require information about the manipulator's behaviors. In our work, in order to achieve natural interactions, we did not use any external sensors or wearable devices for manipulators, therefore, all of necessary information of the scene should be extracted from robot eye cameras. In particular, perception modules should predict movements of manipulator's arm in real-time such as "resting', "grasping a cube", "moving the cube", "end", "none".

For this task, a specific way to solve this problem is using image processing technique (e.g. object detection) to detect a manipulated cube and to track its movements, also assign a **GO** event to an action whenever stroke of movements happen. Some rules should be used to assign a score to an event (e.g. if a cube start moving from a reservoir to the task space, the event can be set as "moving the cube" value, then if the cube stop at the task space, an "end" event is triggered, etc.).

In general, Deep Learning methods can be used to detect human activities [Bac+11]; [Qi+17]; [MR+18]. However, there are still challenges in predicting **GO** events of activities in real-time. Training interactive models directly on image streams may solve this problem. If a lot of data are available, interactive models can be trained end-to-end directly from raw image data (provided by the robot's eye cameras) instead of detecting the primitive actions. For example, a deep learning model combining Convolutional Neural Network (CNN) with LSTM [Don+15]; [Xin+15]) can be use to train end-to-end interactive models in which CNN captures feature of manipulators' activities inside an image frame while LSTM directly captures temporal dependencies between manipulator activities and instructor's interactive modalities.

Another important role of perception modules is **to verify the performance of manipulators.** For example, a perception module should check if a cube placed in a correct position so that it can provide signals for a task model to switch to a new sub-task.

- A task model

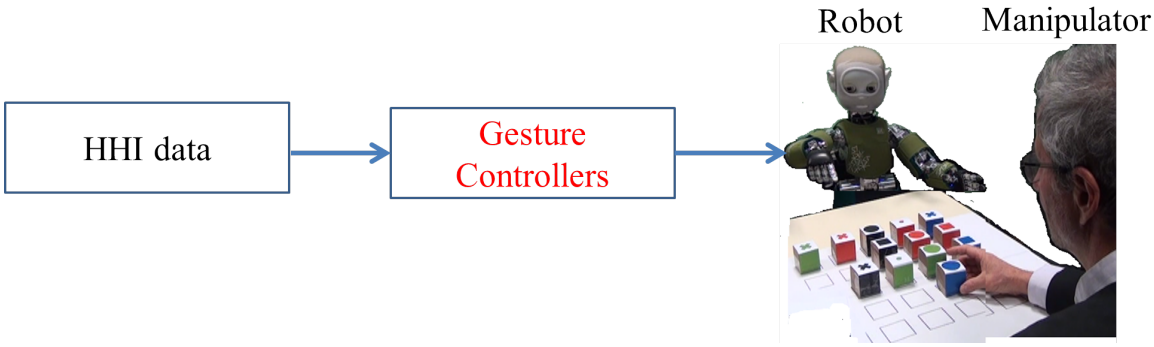
  In our interactive scenarios, there are many repetitive sub-tasks that the robot needs to perform in sequence. A task model can be considered as a planner that manages the order of sub-tasks. Finite State Machine (FSM) can be used to easily describe this task model (see appendix B for examples of a task model for the *RL/RI* scenario).

5.1.2 A strategy to assess/evaluate the modules

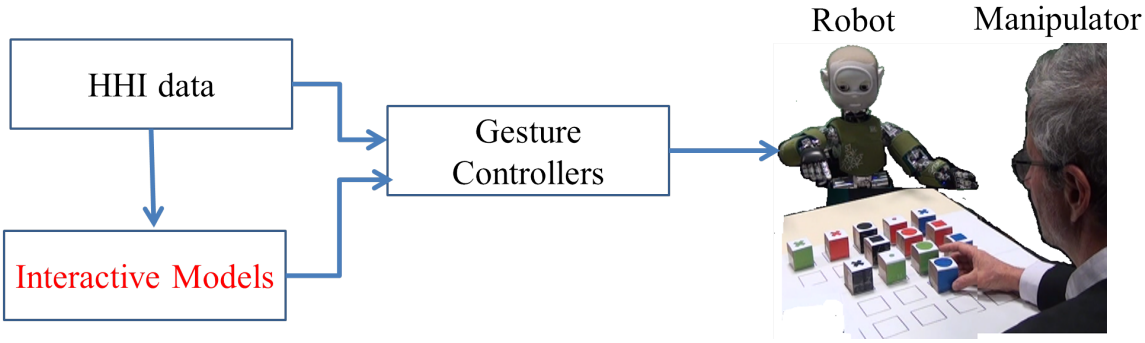
As mentioned before, **each module should be built/verified in both independent and combined ways.** Our evaluation strategy can be splited in two steps:

- Independent Evaluations: each module should be evaluated independently from others.
- Combined Evaluation: all modules will be combined to construct an autonomous robot then they are evaluated in general by letting the autonomous robot interacting with humans.

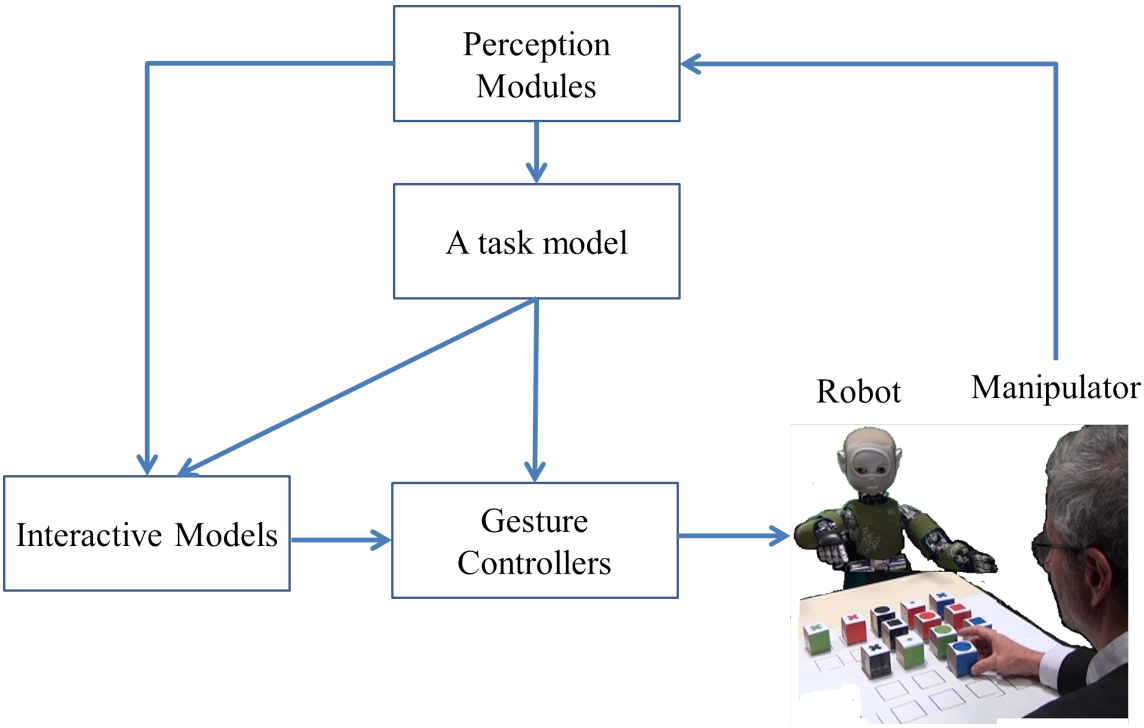
Our strategy to build/verify modules is shown in Figure 5.1. Firstly, in order to evaluate gesture controllers independently with interactive models as well as perception modules, we used directly HHI data to drive robot behaviors as shown in Figure 5.1 (a). Then, we can evaluate (and possibly correct) interactive models generating actions directly on the robot by using HHI data instead of perception modules as depicted in Figure 5.1 (b). Finally, we can evaluate a fully autonomous robot by running all modules (as shown in Figure 5.1 (c)):



(a) Firstly, actions are driven by HHI data, in order to evaluate gesture controllers.



(b) Secondly, Actions are driven by interactive models and HHI data: inputs (perception streams) are provided by ground-truth HHI data in order to evaluate the outputs of the interactive models on the robot.



(c) Finally, perception modules and a task model are integrated too so that the fully autonomous robot can be evaluated

Figure 5.1 – A strategy of evaluating the robot step-by-step: from a replicated version to an autonomous version.

interactive modules receive feedbacks from perception modules to generate actions in order to drive gesture controllers interacting with human subjects.

Robot behaviors are generated in order to interact naturally with humans, therefore, each module (gesture controller, interactive models) should be evaluated by humans. We use here "Wizard-with-Oz" technique ("with" to express that the measurement of Oz is not precise and/or not measured at all [SJS09]) to help evaluate the robot behaviors in the steps. Particularly, the robot here will replicate the scenario where the actions are generated from the interactive data (for evaluating gesture controllers) or interactive models (for evaluating the models). The human here plays as a manipulator and moves the cubes following robot guidance. Here we can use two approaches to evaluate the interactions:

- **Subjective evaluation** In order to detect the "when" and "what" of the behavior errors, we can use our proposed online evaluation framework, as we did to evaluate gesture controllers replicating the *RL/RI* scenario as described in chapter 4.
- **Objective evaluation** We can evaluate the effectiveness of robot behaviors by analyzing how far do manipulators complete the task (e.g. how many cubes are manipulated in right positions, how fast do manipulators respond after the robot gave an instruction, etc.).

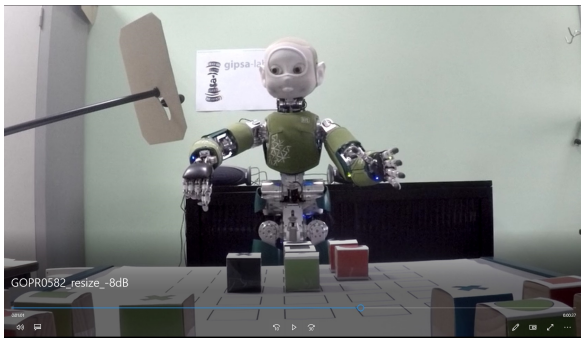
As an example of our strategy, we demonstrate here our robot that instructs a manipulator to move cubes, whose behaviors are generated by interactive models and used to evaluate the models.

5.1.3 The robot replicating the *PTT* scenario

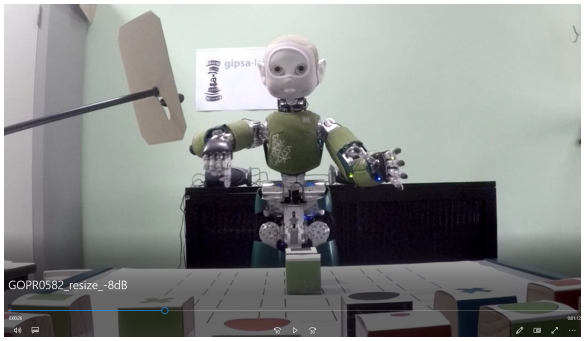
In this section, we describe the robot that can **replicate the *Put That There* scenario** in order to evaluate its interactive models in future. In particular, robot actions are generated by the interactive models, while inputs (perception streams) are feed by HHI data as shown in Figure 5.2 (b). By the way, the interactive models are evaluated independently from perception modules which are not available now as well as independently from gesture controllers which are assumed built and verified before. The interactions are demonstrated in a video following this link[1]. Several robot behaviors are illustrated in Figure 5.2. In the task, the robot first directs its gaze to a tablet to read information of a manipulating cube shown in Figure 5.2(a), then the robot looks at the manipulator face to engage with the manipulator or to confirm he is concentrating on the task (Figure 5.2 (b)). Then, the robot gazes at a reservoir to seek the cube and uses pointing gestures to indicate the cube 5.2 (c)) to be moved to the target position 5.2 (d)).

In the video, we implement the robot replicating the scenario with just arm movements generated by the interactive models, while gaze (drives head movements) and speech are driven

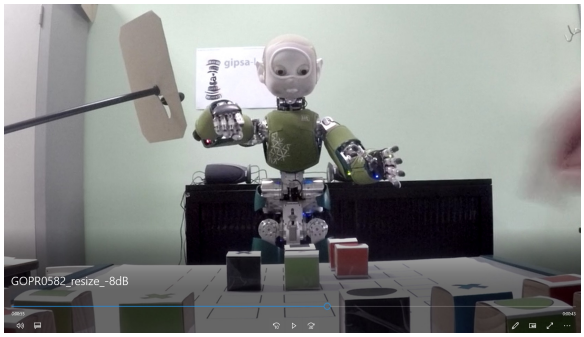
[1] `https://youtu.be/t0CiJaIbJ1w`



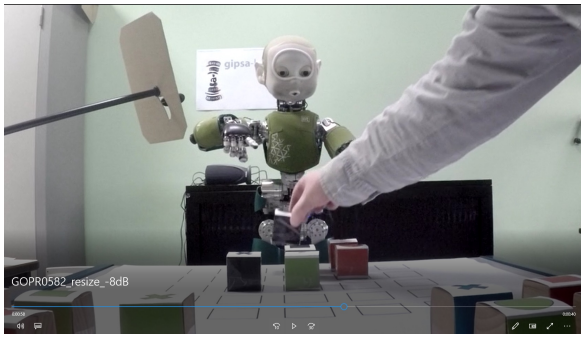
(a) Reading information from a virtual tablet



(b) Looking at the manipulator face



(c) Pointing at a cube in a reservoir



(d) Indicating target position

Figure 5.2 – Excerpts from the PTT replicating scenario, in which arm action events are generated by interactive models, while gaze and speech are driven from HHI interactive data.

by interactive data. However, gaze/ head movements are also available to be driven part-by-part by interactive models. Using the replications (the robot are driven patly by interactive models and partly by interactive data or fully interactive data) can help to evaluate interactive models by analyzing how manipulators' performance is impacted. In particular, the actions of the robot driven fully by interactive data can be used as a "ground truth" so that we can measure how much does he/she finish the task, or compare performance of the robot driven by interactive models to the ground truth.

5.1.4 Comments

In the illustrated video, the manipulator can **finish** the task even when the words of the robot are not much clear. This shows that the interactive models may generate somehow adequate robot actions (e.g. arm movements) to complement speech so that the manipulator can infer correct information that the robot provided.

For the *PTT* scenario, due to the requirements of perception modules, which is out of scope of this thesis, we do not have an autonomous robot to perform the task yet. In the next scenario, we will present the complete realization of a basic autonomous robot that can perform the other task (*RL/RI* scenario).

5.2 A basic (immature) Autonomous Robot performing the *RL/RI* scenario

In this section, we present a basic autonomous control model, which is used as a demonstrator of some of our results and will serve as a baseline for improving robot behaviors in the future.

We focus here on **a basic autonomous control model for the *RL/RI* scenario**, because the perception module basically required in this task just requires a speech recognition for capturing subjects' answers during the interaction. Building real-time perception modules for the robot is out of the scope of this thesis, but speech recognition packages do exist and have be used with success.

5.2.1 Rules-based interactive models for the *RL/RI* scenario

For this scenario, we already built gesture controllers, which were able to reproduce the *RL/RI* scenario by using events extracted from HHI data (see chapter 4). In particular, in order to transfer actions from HHI to HRI, some extra-rules were implemented on the gesture controllers. In this sub-section, we will describe some of the rules and additional rules to enable a robot interact with human subjects.

In fact, the rules designed for gesture controllers provide **intra-coordinations** for robot

behaviors (relationships of endogneous modalities). They are listed below:

- Arm and Gaze The robot should move arm to click to its tablet and mark scores, after gazing at the click-area.
- Neck and eyes Neck and eye movements are driven synchronously by solving inverse kinematic of gazing at 3D points [Ron+16].
- Eyelids and speech

  Eyelids have blinking gestures with frequencies following Gausian distribution $0.5Hz +/- 0.1Hz$. Amplitudes of opening and closing eyelids are coupled with eyes elevation: the higher eye elevation, the more openning eyelids. And eyelids also couple with speech.
- Speech and Gaze The robot should gaze at a subject's face when asking questions.
- Speech Utterances are selected based on uniform distribution of the interviewer's performances. For example, for asking a question in indexed recall, the robot can select randomly one of sentences below:

  1. Est-ce-que vous pouvez me redire le nom du ... ? (Could you retell me the name of ... ?)
  2. Le nom du ... (The name of ...?)
  3. Et alors, ce ...? (Well, the ...?)
  4. C'est quoi? ce ...? (What is ...?)
- Lexicon of Backchannels Backchannels are selected randomly according to the empirical distribution.

In order to interact with human, we added some rules to provide **inter-coordinations** between the robot actions and human patterns. Here are robot actions responding to the human subjects' actions:

- Subjects gives a correct answer The robot mark the score (gazing at tablet and arm moving to click on it) as soon as the subject gives a correct answer.
- Subjects do not answer The robot should give encouragements or reminders to the subject.
- Subjects finish a sub-task The robot speaks to and gazes at the subject face to introduce the next sub-tasks triggered by the dialog FSM.

With these intra- and inter- coordination rules, we have a baseline rule-based interactive model to drive the basic autonomous robot with "naive" behaviors (we call it naive because the robot just performs the tasks without adapting to human subjects).

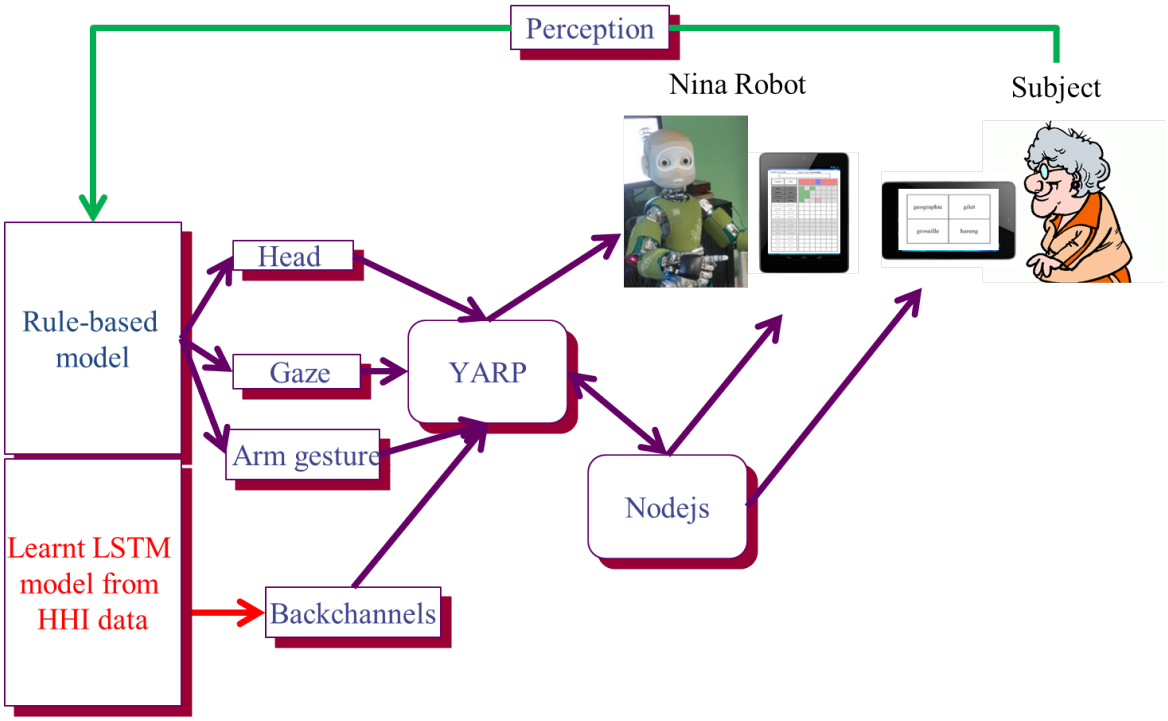


Figure 5.3 – Schematic model of a baseline autonoumous robot interacting with a human subject in the RL/RI scenario. In the framework, the interactive model includes: backchannels generated from a LSTM model and other non-verbal behaviors (head, gaze, arm) generated by rules. The perception module uses a speech recognition system to recognize answers of the human subject. YARP, a middle-ware, provides a protocol to transfer commands from the interactive models to gestures controllers (arm, gaze, head, and backchannels) and to execute the robot behaviors. A Nodejs server is implemented with FSM to manage sub-tasks as well as monitor the displays on the two tablets.

5.2.2 An Autonomous Control Framework

The human being can be seen as a distributed system which is consisted of multiple subsystems working independently but communicating with each other at different scales and levels, e.g. organs, cells, molecules. Inspired by this, a framework for controlling an autonomous humanoid robot is built from modular subsystems, which can reduce *workloads* of the robot's *brain* (here, interactive models).

We design a framework to run the basic autonomous robot as shown in Figure 5.3. The framework includes several main modules: perception of subjects' feedback, interactive models to generate actions (head, arm, gaze, backchannels); and a task model to manage sub-tasks of the scenario.

For implementation, the framework consists of several main components:

- YARP, an open-source midlleware is designed for development of distributed robot con-

trol system [MFN06].

YARP provides intercommunications among subsystems/modules through *Port* mechanism, which manages multiple asynchronous input and output connections between modules. Each module can create many different ports for sending as well as receiving data through network. YARP supports cross-platform compilation and different programing languages, so, easier for us to run different programs/modules on different computers.

Link to human being, our brain controls our body through connections made by the nervous system. The nervous system carries electrical impulses send action signals from our brain to the body. We can imagine that YARP plays the role of the nervous system to transfer signals (here, data bundles).

- Interactive Models

  In our work, interactive models can be consider as a part of *brain* of our robot that perceives the environment and produce actions signals to run the robot body.

  For now, robot action events are generated by interactive models, which are learned from HHI data or constructed by rules as described above. Some modules we provide for the robot to run the *RL/RI* task including: arm and gaze are following described about, head motions are driven by gaze events, eyelids are following speech and elevation of eyes (see chapter 4). Backchannel events are generated from the LSTM-based interactive model (in chapter 3).

- A Task Model

  The robot should be provided a task model in order to manage the state of the interactive scenarios. In the interactive scenario, there are repetitive sub-tasks needed to be run in sequence. In order for the human-robot interactive scenarios to perform smoothly and correctly, we propose to use Finite State Machine (FSM) that are used as a task model to manager sub-tasks that used as a task model for running the autonomous robot. The FSM is implemented on Nodejs and described in detail in appendix (see section B).

- Perception Modules

  Perception modules provide inputs for the interactive models.

  For interacting basically the *RL/RI* scenario, input streams of the interactive model generating backchannels include speech activities of the robots (the robot's speech events are driven by FSM and speech activities of human subjects recognized by Google Cloud Speech API through web-sockets communication protocols using a JavaScript framework (named yarp.js) [Cil17].

A video demonstrating the basic autonomous robot is available via the gipsa-lab website[1]. It shows the robot giving instructions to the human partner, leading the learning step of words 4 by 4, and then validating the answers when they are correct, either in the free recall or in the indexed recall steps. The robot uses the tablet in its hand, clicking to tablet to express making scores, gazing to it or to the user, etc.

[1] `http://www.gipsa-lab.fr/projet/SOMBRERO/videos.html`

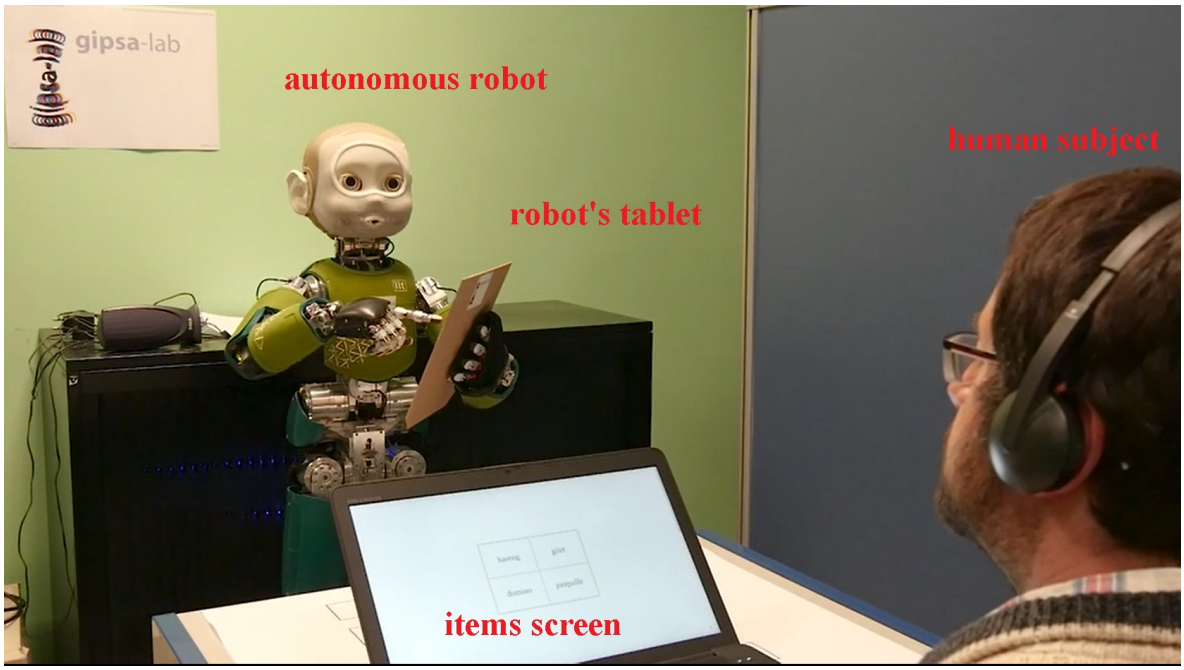


Figure 5.4 – Autonomous robot performing RL/RI task with human subject using rule-based model. Speech recognition uses the Google Cloud Speech API.

5.2.3 Comments

For the *RL/RI* scenario, backchannels of the interviewer often overlap with subjects' speeches. Therefore, in order to emit these backchannel adequately, an incremental speech recognition module should be used to react in real-time to relevant subjects' utterances. For now, we have been used Google Speech Recognition API in synchronous mode which performs recognition on speech audio data sent to the server whenever a subject finish speaking. Therefore, backchannels generated from the LSTM model are seem delayed. In future, we should use a streaming recognition that provides interim results while subject is still speaking. Another solution is to develop a End-of-Utterance (EoU) prediction...

For conceptual design of gazing behaviors, for now, we assume that subjects' face is located in a specific place in 3D space in front of the robot. As a consequence, human subjects can feel robot does not actually look at them. In future, in order to engage with human subjects, the robot should be provided a face detection module to capture locations of subjects' face to gaze at right position.

In fact, the rule-based autonomous robot is just made to interact basically with human subjects. It is not designed to adapt behaviors to different people with different characteristics. In the next section, we describe the challenges and our approaches to step-by-step increase the behaviors and achieve a mature autonomous with higher social level in future.

5.3 Challenges remaining to achieve a widely acceptable autonomous robot

This section overviews these challenges that we still face to get a mature robot with high social levels to interact naturally and adaptively with different subjects. These challenges are:

1. Modeling of **users' profiles:** The robot should adapt its behaviors according to human subjects.

   In order for the robot to be acceptable by a wide range of humans, the robot should adapt its behaviors to different subjects. For example, the robot should change its speech's tone, volume or speech of body movement, etc. in concordance with users' personality and preferences [TM08]).

   A large set of interactive data should be collected to cover statistically significant social factors such as gender, culture, etc. so that the interactive models can capture these aspects. Using such large HHI data to run HRI faces a challenge of scaling from human demonstrators to the robot in term of different perception and action abilities (the corresponding problem). As a consequence, the cost of semi-automatically annotation (see chapter 2) will be really expensive and time-consuming. Therefore, we need more automatic way of annotating interactive data in order to reduce the time-consuming task of hand-crafted annotations.

2. **Transferring actions** from human teacher to the robot: There exists unnatural behaviors due to different durations of robot actions and human actions when adapting scenarios from HHI to HRI.

   In chapter 4, in order to cope with limitations of the robot action abilities, we proposed new scenarios (e.g. using tablets instead of paper sheets). However, there are different durations between robot actions and human actions. Therefore, if we use directly HHI to drive passively autonomous robot, some action events can be delayed. This causes the robot to perform unnatural behaviors. To avoid this problem, we chose adapted actions of robot (see chapter 4) so that their duration time are shorter than those of human actions, but this is just temporal solution. In fact, in order to solve radically this problem, we should collect data directly from human-robot interaction where human is in-the-loop to handle when action events of the robot should be triggered.

3. **Transferring perception:** Scaling from human to robot in term of perceptive abilities

   Although there are breakthrough of Artificial Intelligence (AI) in image, speech recognition, robot's perceptive abilities are still far from human levels. Particularly, when teaching robot, human teacher should be aware if the learner (the robot) could understand/ capture what human are currently doing/saying/looking at or not. This problem requires the learning framework to scale from humans' perceptive abilities to robot perceptive abilities. So, the learning framework should give the teacher feedback on what the robot perceives and what is not seen/perceived/understood. However, it would be quite difficult (or even cognitively impossible) for the human teacher to handle both

the interaction with human subjects and at the same time to track and react to the level of perception achieved by the robot sensors/algorithms. To solve this problem, the learning framework should smoothly enable the human teacher to teach/correct part-by-part of the multimodal modalities so that the human teacher can actually handle the multi-tasking problems.

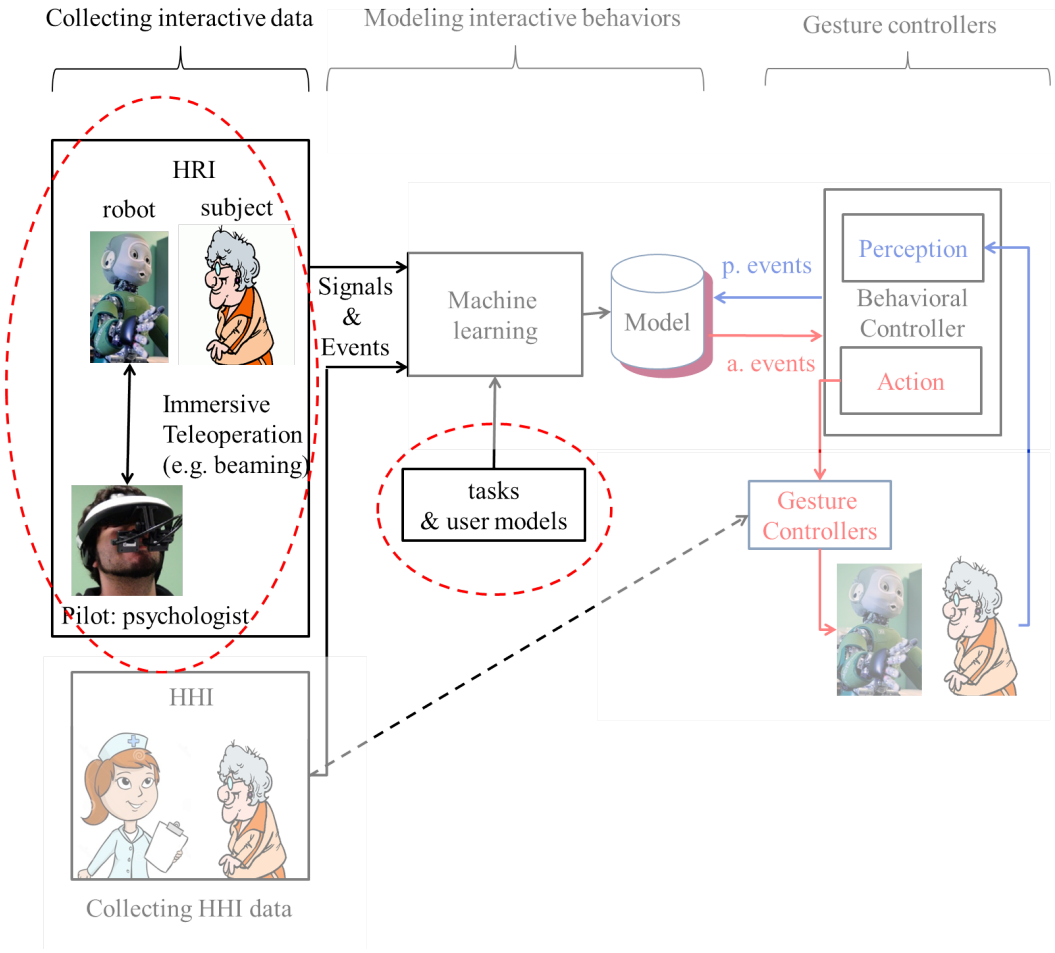


Figure 5.5 – Collecting Human-Robot Interaction data with immersive teleoperation, combined with FSMs. A human pilot picks realtime decisions and actions through the robot body/sensors

The two first problems are mainly related to the interactive data, particularly, how to collect a large amount of data that are comparable with the robot's active abilities and to capture properties of actions corresponding to different users' profiles.

Collecting data from HHI is not straightforward, and complexity their reuse by robots due to corresponding problem. To solve these problems, we propose **to collect interactive data by an immersive tele-operation system** that enables human-to-human interaction naturally through the robot body (robot-mediated interaction). This system allows the robot to observe directly which and when interaction events occur, but also to record how this is

translated into interaction signals: the ones that should be synthesized by an autonomous robot. Therefore, the interactive data, in fact, are collected from the robot's point of view and matches a real, efficient, solution with the sensors/actuators of the involved robot, thanks to embodied demonstration performed by the human pilot, in the same usage context that the planned HRI. This might enable a humanoid robot to **learn faster as well as reduce time-consuming data annotation.** The next section will detail the immersive teleoperation system used to collect interactive data more comparable with the robot abilities.

In the next section, we will present a specific WoZ system: the immersive teleoperation system (called Beaming, developed by Gomez et al [G.G+15] in our Gipsa-lab) and detail how to use the Beaming system for the robot performing the *RL/RI* scenario (see Figure 5.5) in order to collect effective HR interactive data. Also, we propose a method to evaluate the Beaming system before using it to record interactive data.

5.4 Benefits of the Wizard of Oz approach

In fact, there exist various behaviors of human subjects when interacting with social robots: some will be interested by the robot while some feel anxious about them; some may have high expectations on robots because they do not have any preconceptions about the robot capabilities and can be disappointed by faulty behaviors from the robot. Therefore, **humans should be in the loop during the robot training** in order to avoid unexpected behaviors from users caused by the robot limitations/appearance as well as to collect natural data between human and robot interaction: **a model has to be learned in a situation/context matching its future use**. One common technique to enable human remotely operating robots is Wizard-of-Oz (WoZ) technique. The technique refers to a person remotely controlling a robot interacting with its environment such as navigation, verbal, non-verbal behaviors, etc. WoZ enables researcher to simulate capabilities for autonomous robots that are not fully developed or do not exist yet.

There are many benefits of conducting Wizard of Oz in HRI. One of the purposes is to test and evaluate a simulated target robots. Actually, using WoZ is more human-human interaction via a robot body than human-robot interaction. Therefore, the WoZ interaction also could be used to simulate a target robot for measuring acceptability as well as suggesting re-designs. Another purpose of WoZ is to collect interactive data of both Wizard and users. The data can be used to analyze users' behaviors in front of the robot. Also, the data can be used to train multimodal interactive models to achieve an autonomous robot version.

There are several ways that a Wizard can drive robot movements such as joystick for navigation [GHE04], camera capturing key joints for robots' body movements [OKA06], or even drive by kinesthetic driving where the robot is physically guided by human. The WoZ methods are usually used to drive low-level motor skills or tasks where a Wizard forcus on single interactions/skills. For multimodal interactive behaviors, a WoZ system that should not divert the cognitive resources of the wizard from its main task, the targeted interaction with his/her human partner, nor delay the action/reaction times. Typing on a keyboard to let the

robot speak or using a joystick to orientate the head/gaze will not generate the human-like expected reaction times or humanoid head trajectories.

We use here an **immersive teleoperation** system, called Beaming to collect multimodal interactive data. The Beaming stands for **Being in Augmented Multimodal Naturally Networked Gatherings**, which was fisrt introduced by Steed et al [Ste+12], and was developed in our Gipsa-lab by Gomez et al [G.G+15] for an iCub-humanoid robot. With the Beaming system, a Wizard (pilot) is limited to use the robot's own sensors and body to control robot actions. Here, the pilot can drive naturally motions of head, gaze and mouth and speech of the robot to interact with humans as if (s)he were in the robot place. Therefore, the pilot can handle complexed situations when the robot interacts with human subjects.

In this section, we detail the immersive teleoperation system and how to setup the system to collect interactive data of the *RL/RI* scenario. We also describe our plan to evaluate the system.

5.4.1 The Beaming System

As mentioned before, the immersive teleoperation system implemented in our robot is called Beaming technique. The technique is the name of a process that allows a human pilot to transfer him/her-self immediately from one physical place to another body (here, the robot body) in another place. The aim of the Beaming technique is to enable a pilot to feel the strong sensation of ownership regarding his/her body represented by the robot (a.k.a. **Embodiment**). The technique involves two simultaneous parts:

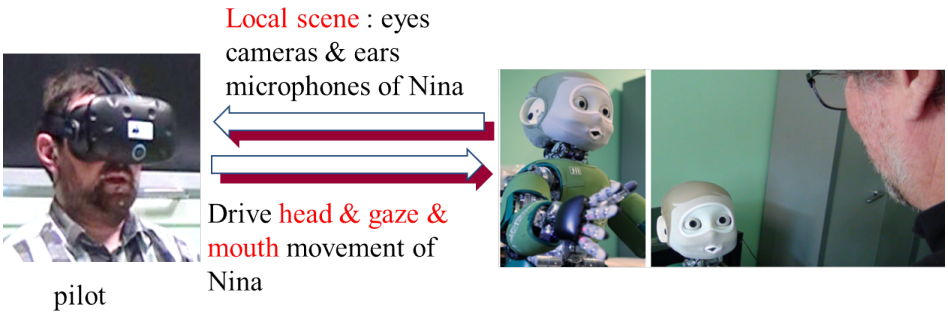
1 **Teleoperation,** to drive robot actions directly by the pilot's actions, physiological states and emotions;

2 **Immersion,** by streaming visual, audio, spatial and context information back from the distant destinations [Gom+15].

In particular, the beaming technique allows the real-time remapping of the pilot's movements to the robot's degrees of freedom so that the robot could mimic the pilot movements. At the same time, the beaming makes it possible for the pilot to sense the scene of the robot local environment thanks to a head-mounted headset worn by the pilot (as seen in Figure 5.6). The pilot can perceive the robot environment through the videos from the cameras embedded in the robot eyes and the stereo sound captured by the microphone inside each robot ear.

An advantage of the immersive teleoperation technique is that it can guarantee actions from the human pilot that can be performed by the robot, which solves immediately the problem of scaling HHI to HRI in term of action skills. For example, during Beaming process, a pilot should drive his head motions in order to adapt to the limited velocity of the robot neck as well as to the limited range of joints.

Figure 5.6 illustrates the Gipsa-lab Beaming system, already described in [G.G+15]. A



(a) Training with joint perception and action streams

(b) Qualysis ® mocap system capturing mouth movement

Figure 5.6 – Immersive teleoperation system used to collect interactive data for the *RL/RI*

head mounted display (HTC Vive) displays views of the robot eyes cameras to the pilot and an eye tracking SensoMotoric Instruments (SMI) is used to track his eye movements. Earphones are used by the pilot to hear sound from robot ears microphones. A Qualysis ® motion capture system captures the pilot's mouth movement to drive robot mouth articulation movements, while a microphone allows to transfer his voice to the speaker located behind the robot's mouth.

5.4.2 Enhancing the system for the *RL/RI* scenario

In the *RL/RI* scenario, a pilot will interact with a human subject for around 20 minutes per test. Therefore, keeping the scenario running smoothly is really important to get natural interaction data between the robot and the human subject. So, the human pilot needs information of sub-tasks of the scenario as well as scores which the human subjects are achieving in order to anticipate actions and give rewards to encourage the human subject. To do that,

we use the "fake" tablet with a virtual-reality marker to display necessary information for the pilot. With an autonomous robot the fake tablet ensures that human subjects "trust" the robot, but with WoZ, the tablet actually provides useful information for the pilot to conduct the scoring as well as to perform more natural interactions. One example of information (16 items words, scores, and state of the scenario), which is displayed on the virtual tablet, is shown in Figure 5.8 (a) & (c).

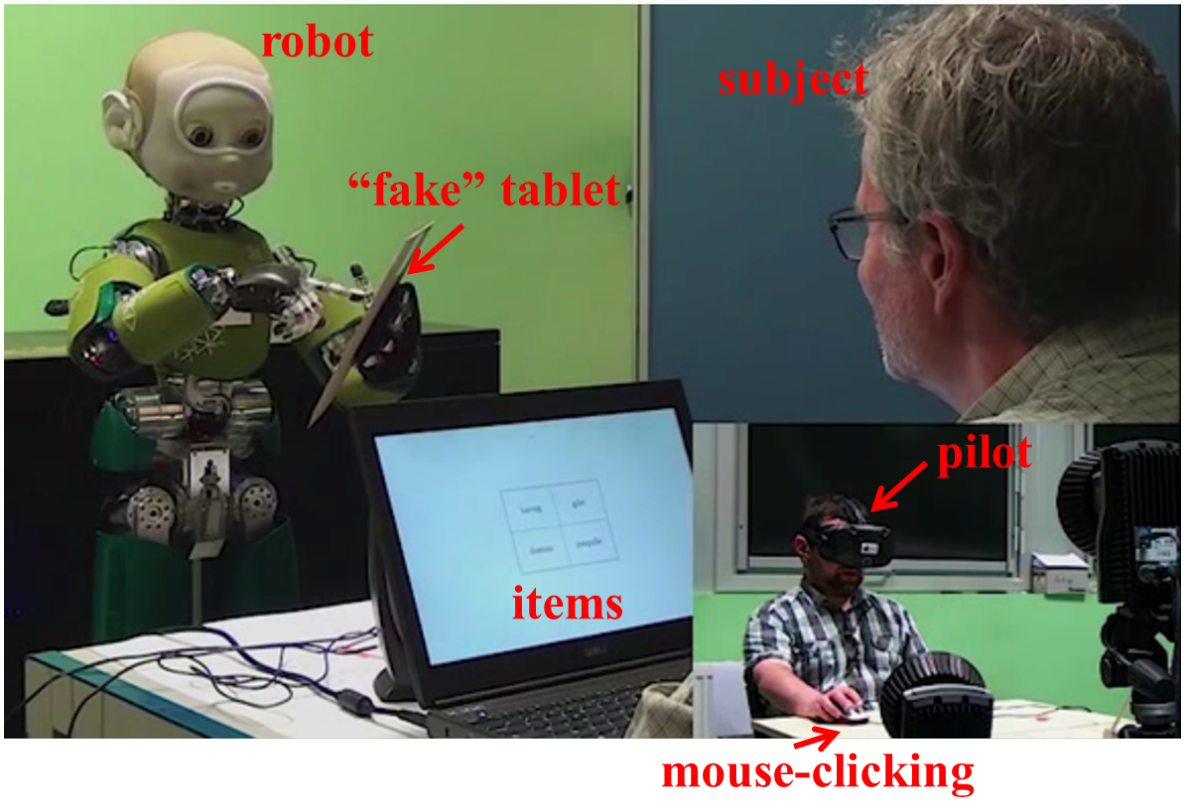


Figure 5.7 – An example of the robot using the Beaming system performs the *RL/RI* scenario. In the picture, the robot is performing its arm clicking behavior (triggered by a mouse-clicking action of a human pilot) on its "fake" tablet when a human subject give a correct answer.

While the Beaming system just enables the pilot to drive the robot head (neck, gaze, mouth and eyelids), we implement here **an additional mouse-control to drive the robot arm with clicking movements.** as shown in Figure 5.7. With the additional arm movements, we expect that the robot will be perceived more natural and engaged by human subjects rather than just using only head movements.

5.4.3 Evaluating the Immersive Teleoperation system

In future, in order to collect the most natural data of the robot interacting with elderly people, we need to verify that our beaming system is adequate. Not only should it be comfortable for

the human pilot (low cognitive load, no dual task effect...), but also we must verify that the pilot behaviors is reproduced smoothly by the robot and **perceived correctly and timely by the users facing the robot.** In this section we propose a simple way to evaluate the system.

There are three methods to measure telepresence systems (e.g. virtual/robotics agents) including subjective measures, behavioral measures and physiological measures [Ins03], which may be used to measure human-robot interactive systems:

- *Subjective evaluations* usually rely on using post-immersion questions. The advantage of the method is that the questions are designed to directly measure the concept we propose to measure. However, the post-immersion questionnaires have the major disadvantage that they do not measure time-varying results and are by nature more biased by events toward the end of the immersion.

- *Behavioral measures* are methods measuring participant feeling by observing behaviors of participant (e.g. participant's movements and postures). These methods can be effected by biases of experimenters (e.g. people, who already used a previous similar system maybe feel familiar than other people who have not any experience on the same systems).

- There are plenty of *physiological techniques* that can be used to measure HRI systems. For example, change in heart rate, with either increasing or decreasing heartbeats per time unit, can be used to analyze some state of a person such as stress, fear, emotion, etc. Change in skin conductance will track the level of conductivity of one's skin and can be used to measure the degree of stress.

Physiological measures are more objective than subjective measures and behavioral measures because they are continuous measure. However, the physiological levels can vary broadly from person to person.

We would like to evaluate the Beaming system in several aspects:

1. comfortable for human pilot,

2. effectiveness of the system,

3. acceptance from human subject.

Because behavioral and physiological measurements are quite expensive to setup experiments, in this work, we propose an evaluation which combines both subjective evaluation using questionnaires and objective evaluation to rate the effectiveness of the immersive teleoperation system.

5.4.3.1 An Evaluation Framework

Here, we design an evaluation framework to evaluate the Beaming system which will be used by a real neuro-psychologist in future to collect interactive data for the *RL/RI* scenario.

The evaluation framework is illustrated in Figure 5.8 (a). Each participant will use the Beaming system to control the gaze and/or head orientation of a robot in an "identify and click" task. They will wear a Virtual Reality headset to see through the robot cameras eyes, and to let the robot potentially imitate their head movements. In their field of view, they can find a screen displaying a word (the target word, as in Figure 5.8) (b), and a virtual tablet with a list of words (shown in 5.8 (c)), including the target word. The task requires participants to look at the screen to read a target word, and then to click on the same word in the list on the virtual tablet display. We want to capture here objectively the time delay between subject's eyes capturing an item in the item screen and clicking event at the item in the virtual tablet. We expect that the shorter delays, the more comfortable the participant as well as more effective system may be. For rating acceptable aspect, we design questions for the participants.

Each participant will perform three conditions of the beaming:

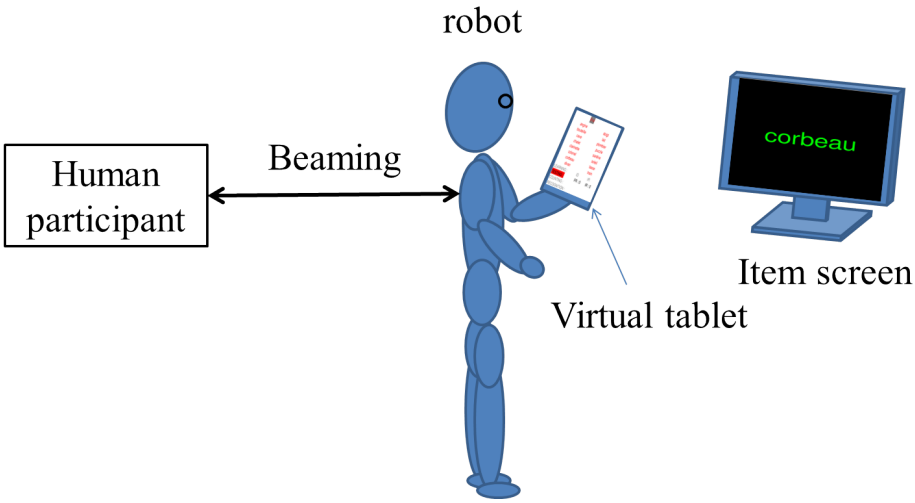
- Fixed Head & moving gaze
- Fixed Gaze & moving head orientation
- Both movable: head & gaze

5.4.3.2 Item selection

In each condition, 16 selected items will be tested in two rounds. The distance between two consecutive testing items will follow uniform distribution which cover all of distance in the table of items as shown in Figure 5.9. The items are sorted and merged to be balanced in order and avoid that two neighbor items are in consecutive tests, as shown in Figure 5.10.

5.4.3.3 Illustrative Examples of Evaluating Data

1Figure 5.11 illustrates eyes and head movements of a participant performing the clicking-beaming evaluation in the three conditions. In the fixed head condition, the gazes move with higher amplitude (gaze-1 and gaze-2) and the time for finishing the test is the largest one from the three conditions due to the large saccade errors. In the fixed robot eyes condition, the head just seems to move to one position (see Figure 5.11 (b) so that the participant can see the items from the 2 sources (tablet and desktop screen) in the head-mounted display screens. At that time, he tries to do his best to just move his eyes to see the displayed item and click. Therefore, the duration time for this test become shortest. Finally, with movable both of eyes-head conditions, the participant move both head and eye at the same time. Because



(a) Schematic modal of a beaming evaluation

(b) A target word, as seen on the screen

(c) The list of words on the virtual tablet, where the target word should be found and clicked

Figure 5.8 – Beaming evaluation using a virtual tablet (c) and a screen for items (b)

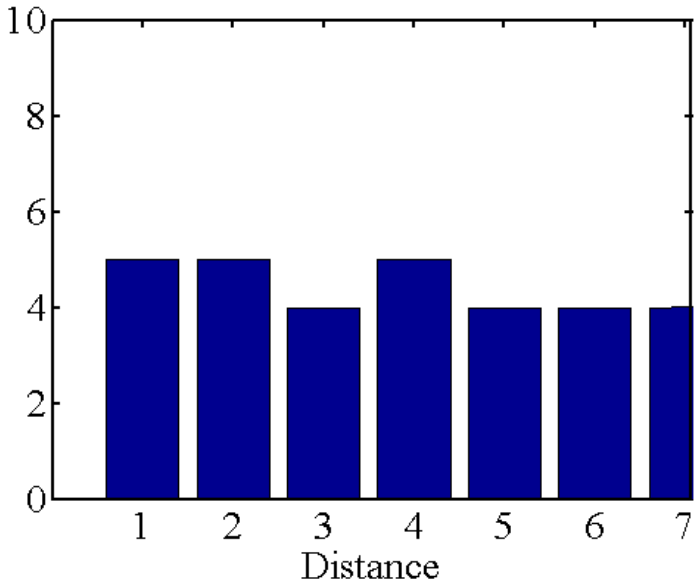


Figure 5.9 – Distributed distance between two consecutive items

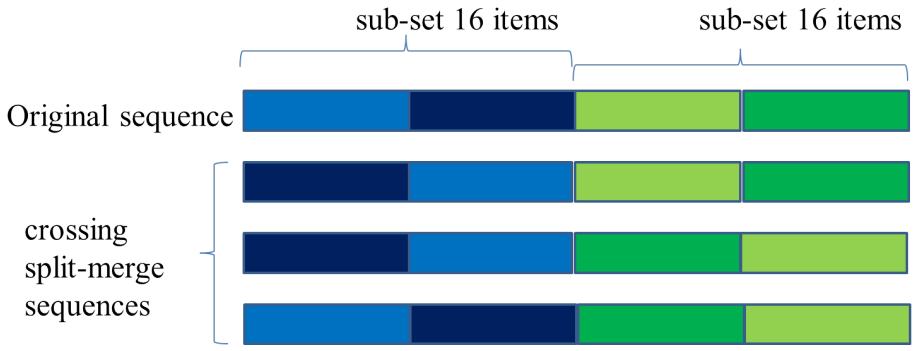


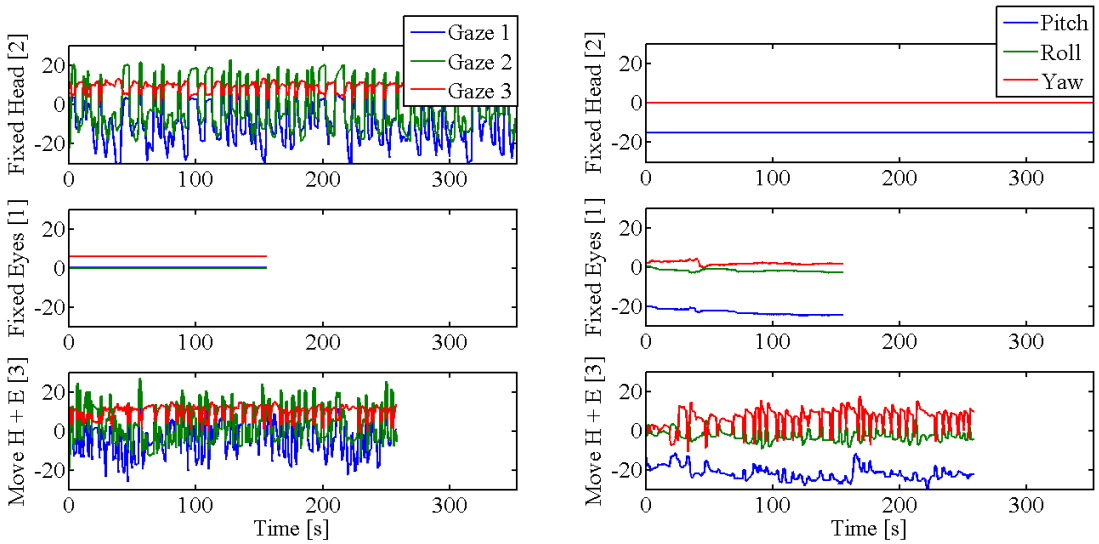
Figure 5.10 – The items are sorted and merged to exhibit a balanced order

of latency between the participant eyes movement and the display from robot camera, the participant tend to move slower his/her eyes so that the head will contribute more largely in gazing targets as is the case when human gaze with their own head/eyes.

The evaluation system is ready to perform: in near future, we will have more data to analyze and validate the Beaming system performances before its use to collect interactive data.

5.4.4 Comments

The evaluation framework focused on **pilot's point-of-view**, which measures the effective of the Beaming system. In future, we also can measure acceptance of the Beaming system in



(a) Eye movements (b) Head movements

Figure 5.11 – Gaze and head movement of a subject in clicking-beaming evaluation

subject's point-of-view, in which participants interacting directly with the robot controlled by a professional human pilot.

5.5 Conclusion

In this chapter, we discuss our perspective on how to build autonomous robots that can perform the short-term interactions. For now, the robot has been able to **replicate the *PTT* scenario,** in which actions are generated from interactive models as well as from interactive data. We discuss our strategy to build perception modules and evaluate required modules (gesture controllers and interactive models) so that the robot can collaborate with human naturally and effectively in future.

For *RL/RI* scenario, a basic autonomous robot is built with rule-based models to drive gaze, arm, speech while backchannels are generated by interactive models; and a speech recognition system is used to provide inputs for the interactive models. We also presented the challenges we are facing to achieve a more mature autonomous system with higher social levels.

Most of the challenges are concerned by interactive data. In order to progress against these limitations, we propose to apply immersive teleoperation system, which are integrated with Finite State Machine to collect naturally interactive data for HRI.

The immersive teleoperation system allows a pilot (e.g. a neuro-psychologist) to interact with human subject through the robot body. In fact, the human pilot can perceive the scene

through robot eye cameras and ear microphones. At the same time, (s)he can drive the robot verbal and nonverbal behaviors such as head, gaze and arm movement. Therefore, the pilot can control robot to adapt its behaviors when interacting with different subjects. We also present how to evaluate the immersive teleoperation system before using it to collect the interactive data.

In fact, with the immersive teleoperation system, the interactive data could be more natural for teaching the robot how to perform social interaction, with more variability. However, it still has difficulties in scaling from human perception abilities to the robot abilities. The pilot would have difficulties at the same time interacting with human subjects and care what the robot can learn. We will discuss a semi-autonomous robot which shares control between human pilot and the robot in the next chapter. Therefore it can reduce loading of the human pilot and enable the pilot to focus on the robot perception limitations.

Chapter 6

Conclusions and Perspectives

6.1 Conclusions

This thesis main concern/goal is about a humanoid robot that can interact naturally with humans through verbal and co-verbal behaviors. We propose here **a learning framework** that enables a humanoid robot to learn **multimodal interactive behaviors** from human coaches, which we detail and use with 2 short-term interactions and dedicated behavior models, with a final demo implementation on the iCub robot.

The learning framework includes three main parts: (1) collecting interactive data of multiple modalities (speech, head, arm, gaze, etc.) from human-human interaction (HHI) and process the data to get events and features that can be scaled to HRI; then (2) build multimodal interactive behavioral models, which can capture both inter- vs. intra-coordinations between the modalities of interactions (e.g. speech, arm, gaze, head motions, etc.) to generate actions from perception streams; finally (3) make gesture controllers to execute actions generated from the models.

We focused on short-term interactions in order to collect more easily enough interactive data to train interactive models. We selected here **two short-term interactive scenarios**, which our robot will be involved:

Collaborative performative task The first scenario is "Put That There" (PTT), in which the robot will play the role of an instructor who instructs manipulator to move cubes. This scenario requires the robot to perform precise coordination between speech and other non-verbal behaviors (arm pointing, gaze attention, etc.). The intra-coordinations concerns the time relations between triggered events of the robot's speech and its gaze and arm movements. In contrast, the inter-personal coordinations concern the robot's sensitivity to the actions of its human partner (here, manipulator's arm movements). By these multimodal interactions, the manipulator is expected to better perceive the robot's instructions than ones given by unimodal interaction (e.g. just by speech). Therefore, the collaborative task can be performed more efficiently.

Interview The second task is a neuro-psychological test, namely a *Selective Reminding Test* (French RL/RI). In this task, the robot plays the role of a neuro-psychologist who interviews elderly people to diagnose potential Alzheimer disease. This task requires the robot to perform not only adequate verbal and nonverbal coordinations but also mimic

professional skills of the psychologist to keep engagement of the subjects and encourage them to foster their memory. In particular, the right choice and timing of backchannels play an important role.

Multimodal data collected during HHI experiments are then processed to get useful features for training multimodal interactive behavioral models. We compared performance of Long-Short Term Memory (LSTM), a Deep Learning method, with that of statistical models, previously developped in the laboratory, to construct interactive models. The LSTM-based techniques are used **to generate discrete events** (gaze, arm, interaction units for the *PTT* scenario, backchannels for the *RL/RI* scenario) as well as continuous variables (head motions for the *PTT* scenario). Compared with the statistical methods, **LSTM generates not only better prediction performance but also provides better coordinations between multimodal streams** (measured by Coordination Histogram). These good results are explained by the ability of LSTM to capture long-term time dependencies, which provide decision layers with relevant contextual information. **For head motion generation,** we proposed a cascaded LSTM model, which first uses one LSTM layer to predict gaze as an intermediate task and a second LSTM to generate head motion. By this way, we explicitly provide a causal relation between head and gaze.

We then **build gesture controllers** that are used to execute action events generated from the interactive models. We also **propose an online evaluation framework to detect faulty behaviors** of gesture controllers, and helps improving the models. The online evaluation requires participants to continuously watch a video from the viewpoint of a subject interacting with the robot and just press a key (ENTER key) to signal faulty behaviors. Time distribution of these "yuck" responses exhibits clear maxima that cue elementary faulty behaviors that system devloppers can then handle. We evaluated the gesture controllers with action events directly driven by HHI data in two versions of the robot: (1) an initial version; (2) a corrected version that benefits from the correction of faulty behaviors identified bythe first assesment. We found that the corrected version effectively reduces the number of yuck responses. This **confirms the effectiveness of the framework for detecting the robot's faulty behaviors.** By analyzing the post-hoc questionnaires and the free comments given at the end of the test, **we found that people seem to have higher expectations when the robot is less faulty.** For example, in the second experiment, participants usually comment "why does the robot not smile?" or "it is better if the robot give a joke", which did not usually happen in the first experiment. In future, the same procedure can be used to actually rate the autonomous robot with fully functional interactive models.

6.2 Perspectives

In order to build interactive behavioral models for the autonomous robot which would cope with the different users' profiles, our learning framework faces many challenges. In this section, we discuss some of them.

6.2.1 Evaluation Framework

For now, while watching the video of robot behaviors, the participants just press a key to signal the faulty behaviors and give their comments at the end of the video. Participants may not remember all of faulty behaviors for giving comments. Therefore, the inverse engineering step to annotate the faulty behaviors cannot be performed automatically. In the future, the evaluation framework can be improved by enabling participants to signal the faulty behaviors and give comments *on-the-fly* by selecting some options – e.g. select different keys – or *pause-to-comment*. In this way, we can annotate the faulty behaviors automatically or semi-automatically without the bias induced by inverse engineering.

6.2.2 Collecting and Annotating Interactive Data

Collecting and processing data are important steps of our learning framework. These operations as well the scaling from HHI to HRI are time-consuming works. In our last chapter, we proposed to use our new immersive teleoperation system which enables a pilot to interact with a human subject through the robot body. By shaping the pilot's action of humans to the sensorimotor abilities of the robot, we can collect directly suitable features for training interactive models. Therefore, the time-consuming annotations and adaptation are significantly reduced.

6.2.3 Improving and Adapting Robots' Behaviors

In this section, we will describe our perspective on how to improve and adapt our robot behaviors in the future.

6.2.3.1 Sharing Control

When a pilot teaches the robot his/her interactive behaviors, in fact, he/she should aware of what the robot is able to perceive so that the pilot can provide adequate actions for the robot. With full teleoperation, the pilot has difficulties in both handling natural interactions with the human subjects and caring about robot perceptive abilities.

In order to solve this problem, it could be interesting to transfer smoothly perceptive abilities from the human pilot to the robot. In this approach, the robot will take control over the pilot part-by-part, starting with low-level autonomous levels to higher-level behaviors. This technique can be found in literature as semi-autonomous teleoperation or shared autonomy, or assisted teleoperation [Mas+15]; [HC18]; [Li+15]. This continuous adaptation of shared control should enable the pilot to progressively concentrate on high-level cognitive abilities and decision making, while the robot takes in charge the low-level sensorimotor abilities and reactive behaviors. By the way, the robot will increase knowledge and abilities, therefore, the

number of necessary teleoperator interventions will decrease over time [Mas+15]; [Mas+10]; [Shi+15].

6.2.3.2 Generating variably interactive behaviors

In fact, human behaviors (e.g. head motions) are quite various, complex and stochastic in nature. However, for now, due to limited interactive data collected, we build here interactive models that generate robot behaviors in a deterministic way (e.g. deterministic output of discrete events/ continuous values). In future, in order to generate robot behaviors with more variability, the interactive models should generate outputs in stochastic ways instead of deterministic outputs. For example, an approach might combine LSTM and Gaussian Mixture Model (GMM), in which LSTM will predict parameters of GMM (e.g. means, covariances), then, the GMM is used to generate randomly outputs (as is the case of handwriting generation in [Gra13])

6.2.3.3 Adapting the Interactive Models: Transfer Learning

When we release the robots to the world, robots should be able to learn new behaviors by themself or update pre-trained interactive models in order to improve behavioral qualities. How to make robots flexible enough to learn new skills and adapt to new environments (interacting with new subjects) so that the robot can perform open-ended learning?

In fact, collecting a lot of real-world data to train robots performing multimodal interactions is expensive and time-consuming. Therefore, robots should be able to learn new skills that require just a few of new interactive data to perform new tasks based on prior knowledge learned before. Transfer learning (TL) refer to techniques that enable a machine to learn a new task (called target task) faster from an already learned task (called source task) (see a survey of transfer learning in [PY+10]). Therefore, TL is a good choice to reduce the burden of the collecting data. We discuss here several potential applications of transfer learning to provide new skills for robots:

- **Transferring tasks:** Starting from a task already learned, a robot using TL can learn new similar tasks faster [Ros14]. This is like human beings, e.g. one, who can play guitar, is able to learn playing piano faster than another who does not have any skills in musical instruments.

- **Transferring styles:** In order to be accepted by different human subjects, the robot should be able to change its behaviors according to different user profiles. A concept of style embedding [Wan+18] could be used to encode different types of interactive behaviors so that the robot can adjust its behaviors according to different human subjects.

- **Sharing knowledge between different robots:** For now, we just teach behaviors for only one robot. However, if robots can share sensorimotor experiences and cognitive

abilities to each other, this will foster learning new skills and enable robots to adapt with different human environments. TL is not only to help a robot learn new skills from its own experiments but also can learn experiences from other robots [HS17]. This is promising technique to enable robots to share knowledges to each other in order to learn new skills faster to adapt with unstructured human environments.

- **Atomic interactions:** A very challenging issue is here to be able to segment the observed or demonstrated behaviors into elementary skills that can be scaled from existing ones or that are new and require explicit training. The concept of pragmatic frames proposed by Vollmer et al [Vol+16] is quite similar to our interaction units and may constitute an interesting bootstrapping framework.

Appendix A

Sensorimotor Calibration for Pointing

A humanoid robot has many degrees of freedom (DoF) with many sensors and motors as well as other mechanical elements. Therefore, the limited accuracy of sensors or errors from misalignment among mechanical parts can result in imprecise robot behaviors. In this appendix, we present a sensorimotor calibration procedure that links gaze to index pointing. While based on the inverse kinematics, it ensures that the robot actually points its hand/finger more precisely to the object's location to which it is gazing at (the robot points to where it looks).

There are three main approaches for sensorimotor calibration: model-based, model-free and hybrid. Model-based methods require mathematical models of the robot kinematics as well as models of the robot's eye cameras. These methods demand complex calculation and precise calibrations of cameras and other sensors (e.g. position, velocity). In contrast, model-free approaches use machine learning to map directly perception cues to motor commands (e.g. from pixels to arm positions), but their precision depends on an extensive training set. Hybrid methods combine both of above methods: a baseline uses mathematical models (e.g. kinematic models) that is corrected by machine learning methods (e.g. neural networks).

There are also several ways to make the robot's arm point to the object location where the eyes are gazing at (see Figure A.1). In our work, gaze-arm mapping (arrow (2)) for pointing is performed by an hybrid method so that we can use available modules of our robot. More precisely, the built-in Cartesian Controller module [Pat+10] provides the mathematical model of the robot arm's kinematics as well as the inverse-kinematics model (mapping (5)). The module computes commands for the robot' arm to perform pointing under a set of constraints – here straight line relating its shoulder, index finger and the pointed object – so that we can easier collect a large of training data. The iKinGazeController [Ron+16] module provides invert kinematics of the robot head (neck & eyes) as well as mapping (3) between 3D points and pixels coordinates of the robot eye cameras. Based on the iKinGazeController, we built a program called **stereo-click** that instructs the robot to gaze at a 3D position in space given the pixel coordinates of the desired location on the two eye cameras.

We used a red laser pointer attached on the robot index finger to spot where the robot is effectively pointing at and further use the difference between the red spot on the eye cameras and the desired target.

The Cartesian controller is first used to drive robot's arm to point at positions on a chess board in front of the robot. Once the movement is performed, we collect the effective pointed spot using the stereo-click program. We collect the position and orientation (x_d, o_d)

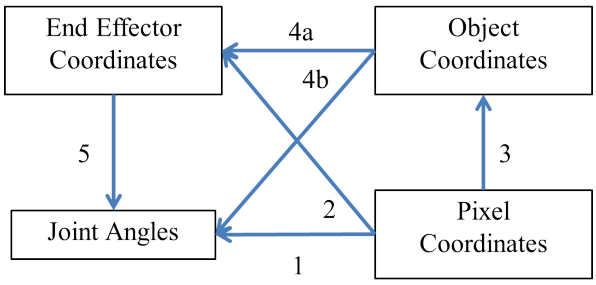
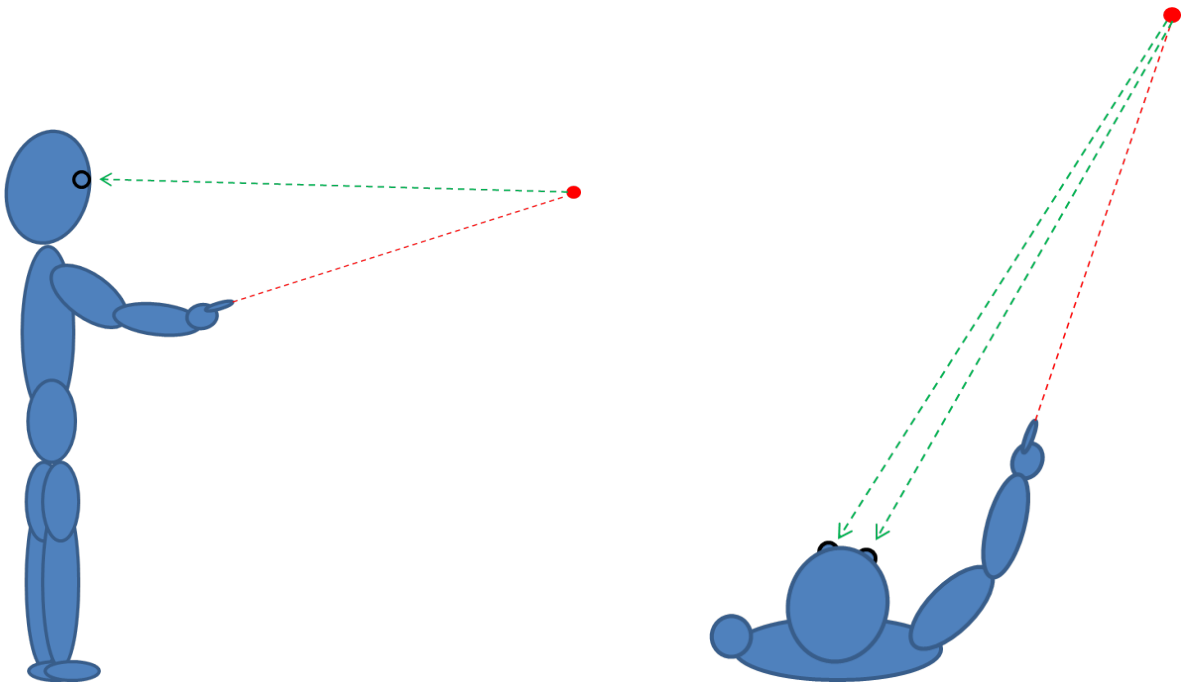


Figure A.1 – Different possible approaches for mapping pointing gestures [Lem+13].

of end effector and 3D positions (x_l) provided by the robot eye cameras using the pixel-to-3D mapping (3). Finally, we use a nonlinear method (e.g neural network) to map (x_l) and (x_d, o_d) (mapping 4a). This mapping to directly compute arm and wrist motions for pointing an exact 3D position by selecting pixels from the robot eye cameras (following mapping 3 and 4a).



(a) A horizontal view (b) A vertical view

Figure A.2 – Schematic model of arm pointing with laser spotter attached on the index finger.

A.1 Comments

In this work, we perform a sensorimotor calibration of arm-gaze pointing gestures. The main purpose of this calibration is to enable a better coordination between gaze and arm to perform the *PTT* scenario, in which cubes are placed on a table in front of the robot. Therefore, the

data that we collected here are positions of laser reflection points in a chessboard attached on a table. The same strategy can be used to calibrate the pointing gesture in an extended 3D space by placing targets in different positions in the sensorimotor working space of the robot. This training could be performed incrementally and incorporate more DoF such as head and torso movements.

Appendix B

Finite State Machine of the RL/RI scenario

In the interactive scenarios, there are many repetitive sub-tasks and sequences of sub-tasks that robots have to perform. In order the HRI scenarios to perform smoothly and correctly, we used Finite State Machine (FSM) models: they describe the dialog as a series of conditioned interactive sub-tasks.

In this appendix, we introduce FSM and their application to the modeling of the sub-tasks of the *RL/RI* scenario.

B.1 FSM

FSM is an abstract machine that can be in exactly one of a finite number of states at any given time. The FSM can change from one state to another in response to some external inputs; the change from one state to another is called a transition.

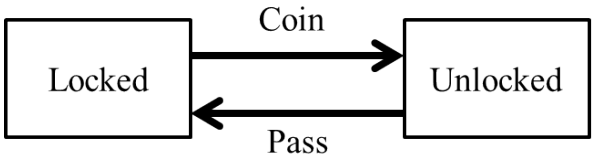


Figure B.1 – A simple FSM model of a Turnstile

An example of a simple FSM is a turnstile which is used in a subway station to rotate gate with a coin receiver shown in figure B.1. At the beginning, the machine is in the (locked) state which does not allow to pass through. When a passenger inserts a coin, the machine switches to the (unlocked) state that allows passing the gate. However, this simple FSM model does not explicit all the turntile's behaviors. For example, what should happen when someone passes through the gate without depositing a coin? Or what should happen if a coin is deposited before the turnstile is locked. In both cases, the state of the gate should be inquired. The final transitions of the detailed turnstile is shown in figure B.2. A system could be modeled by a FSM if it satisfies several conditions:

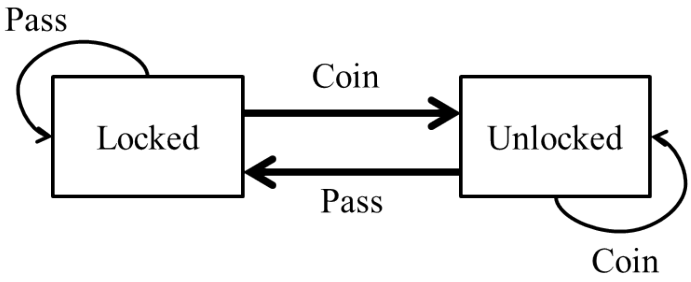


Figure B.2 – A more detailed FSM model of a Turnstile

- The number of states of the system should be finite
- There are a set of finite events that trigger transitions between states
- The behavior of the system only depends on the current state and the current event triggering that state.
- The system has an initial state

Each condition in the FSM corresponds to a logical state transition in the machine. At a particular moment, the machine is only in one state and the state of machine will only be changed to other state when an event is triggered.

B.2 FSM models of the sub-tasks of the RL/RI scenario

We built an FSM of the RL/RI scenario which is used by the robot to control the sub-tasks of the scenario that includes 4 main phases: learning, testing, recognition and counting. The FSM thus includes several sub-FSMs corresponding to each phase of the RL/RI scenario.

The sub-FSM of the *learning* phase is shown in Figure B.3. From an **initial** state, when a “select group” event is triggered, the machine will change to the **Selected Group** state. There, if a “select ID" event is triggered, the machine will switch to **Identifying Items** state. At this state, some actions are to be executed: displaying the 4 items, scores made available to the robot – or its pilot when teleoperated – so that the scoring sheet is updated when the subject’s answers are correct, etc. Then it goes to **Immediate Recall** state when a “select IRC" event occurs: ready for the sub-task of the *learning* phase. When all items have been found, a “finish IRC” event is triggered. If the 4th group is finished, a "finish final group" event is fired immediately and the FSM will turn to the **Ready Counting Phase** state that is the initial state of the next sub-task FSM.

The FSM of *Counting* sub-task is shown Figure B.4. It has 4 states: **Ready Counting task**, **Counting**, **End Count** and **Done Count**. When the state of the machine is **Ready Counting task** and a “start count” even is triggered, the state is changed to **Counting**.

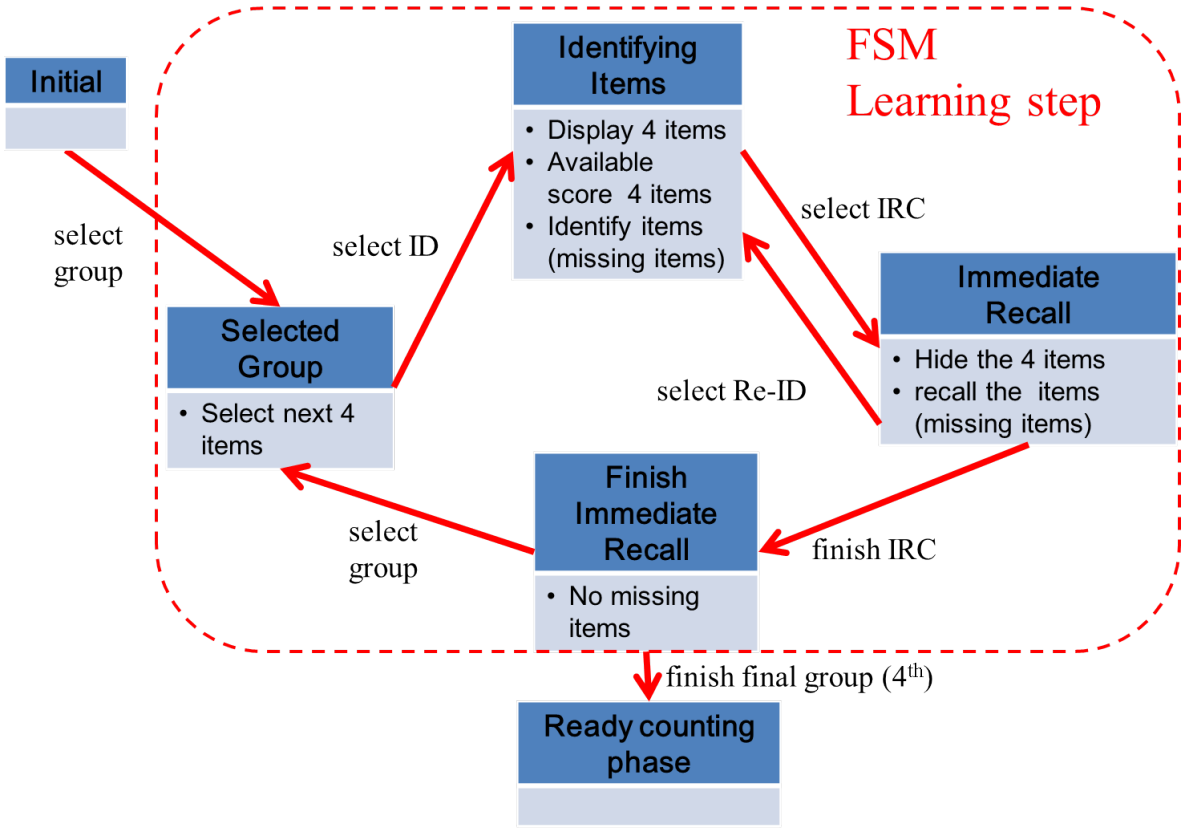


Figure B.3 – A sub-FSM modeling the learning phase

From the **Counting** state, if a "Stop Count" event is triggered, the state will be changed to **End Count**. From here, if a "count again" event is fired, the state will change back to **Ready Counting**. This allows the counting task to be performed many times. Otherwise, if a "done count" event occurs, the state will change to **Done Count** state. This is the end of the counting task and the state is ready for the *testing* phase.

The *testing* phase includes 2 states, which is **Free Recall** and **Relearn**. When the machine state is **Free Recall** and a "Re-learn" event is triggered, the state will turn to **Relearn**. From here, if a "Finish re-learn" event is triggered, the state will turn back to **Ready Counting task**. After visiting 3 times the **Free Recall** and **Relearn** states, the machine will change to **Ready Recognition step** state. It's ready for the final phase of the scenario.

The final sub-FSM is designed for *recognition* phase as illustrated in Figure B.5. From the state **Ready Recognition step**, if a "show word" event is triggered, the state will turn to **Show word** state. At this time, if a "make score" event is fired, the state will turn to a **Subject answered** state. From this state, if "show word" event is triggered, the machine will come-back to the **"Show word"** state. This is performed repeatedly until 48 words are tested. An event "Finish 48 words" is then triggered and the system switches to **End Test** state.

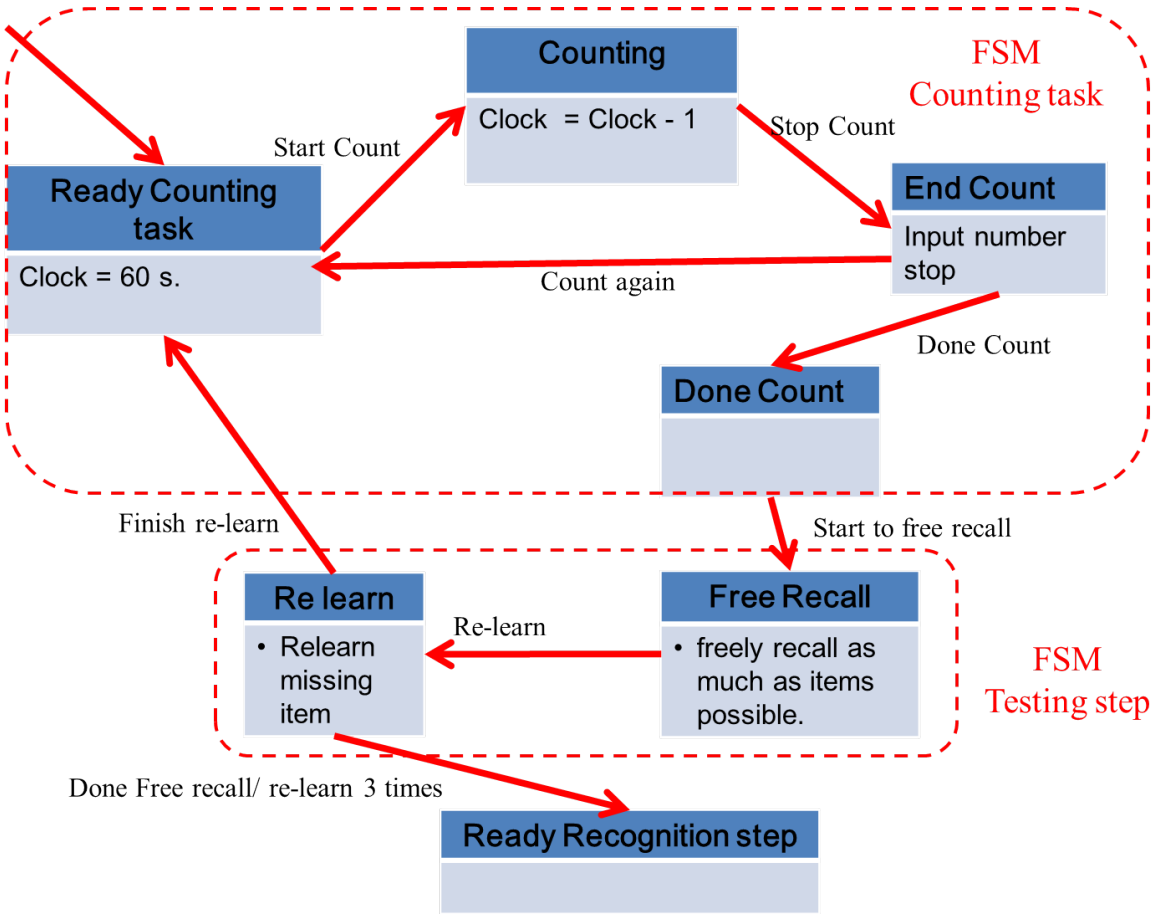


Figure B.4 – Sub-FSMs of counting and testing phase

B.3 Comments

The Finite State Machine (FSM) module was developed to manage the sub-tasks of interactive scenarios, which can be used for both autonomous robots and WoZ systems. It not only keeps the interaction running smoothly by informing the pilot about the state of the scenario, but also compensates for the limitation of the Beaming system such as controlling the arm movements. For example, in the *RL/RI* scenario, when the pilot wants to mark scores for the subject's answer, he/she just clicks on a virtual tablet (where the state information is displayed) and the system takes in charge the sets of actions to be performed by the robot, such as selecting the right position on the robot's tablet to click with the arm gesture controller. Furthermore, in the autonomous version, the gaze is also driven to anticipate the targeted clicking area. Thus, the FSM is used as a *hybrid interaction strategy controller* [Seq+16], playing the role of a planner to drive autonomous robots.

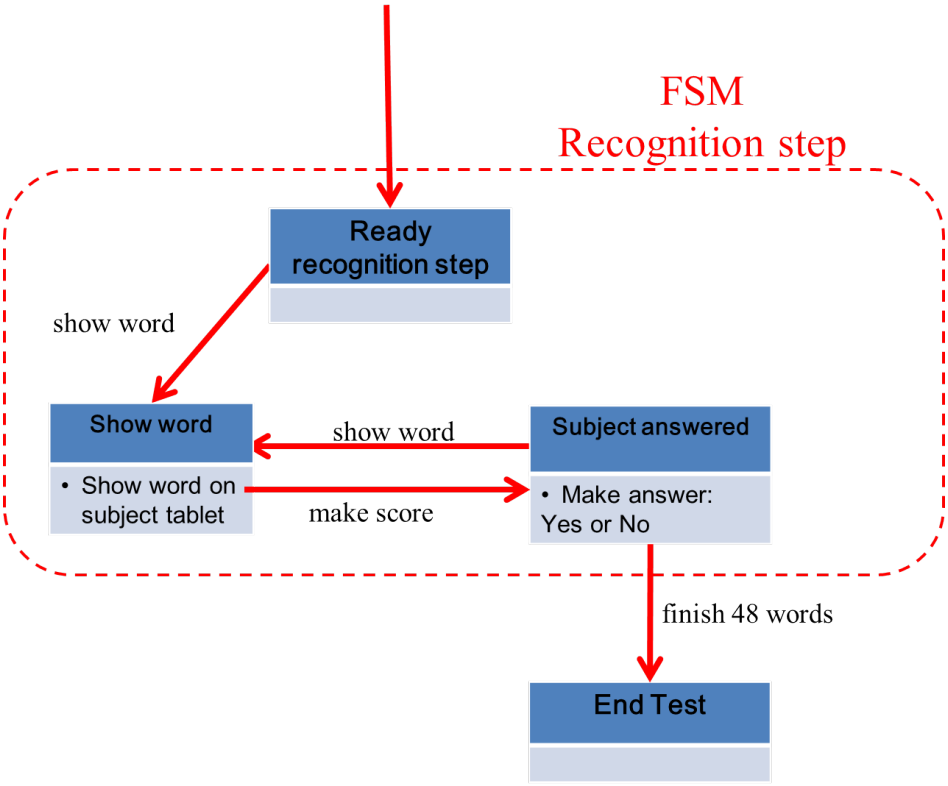


Figure B.5 – A sub-FSM of recognition phase

List of publications

- Nguyen, D.-C., G. Bailly & F. Elisei (2018) Comparing cascaded LSTM architectures for generating gaze-aware head motion from speech in HAI task-oriented dialogs, HCI International, Las Vegas, USA: pp.164-175.
- Nguyen, D.-C., G. Bailly & F. Elisei (2017) "Learning Off-line vs. On-line Models of Interactive Multimodal Behaviors with Recurrent Neural Networks", Pattern Recognition Letters, v. 100, pp. 29-36.
- Nguyen, D.-C., G. Bailly & F. Elisei (2017) An evaluation framework to assess and correct the multimodal behavior of a humanoid robot in human-robot interaction, Gesture in Interaction (GESPIN), Posznan, Poland: pp. 56-62.
- Nguyen, D.-C., G. Bailly & F. Elisei (2016) Conducting neuropsychological tests with a humanoid robot: design and evaluation. IEEE Int. Conf. on Cognitive Infocommunications (CogInfoCom), Wroclaw, Poland, pp. 337-342.
- Nguyen, D.-C., F. Elisei & G. Bailly (2016). Demonstrating to a humanoid robot how to conduct neuropsychological tests. Journées Nationales de Robotique Humanoïde (JNRH), Toulouse, France: pp.10-12.
- Nguyen, Q. V., L. Girin, G. Bailly , F. Elisei & D.-C. Nguyen (2018)(accepted) Autonomous Sensorimotor Learning for Sound Source Localization by a Humanoid Robot, 2nd Workshop on Crossmodal Learning for Intelligent Robotics

Bibliography

[AB93] Mamoun Alissali and Gerard Bailly. "COMPOST: a client-server model for applications using text-to-speech systems". In: *Third European Conference on Speech Communication and Technology*. 1993 (cit. on p. 85).

[Abd+18] Jordan Abdi et al. "Scoping review on the use of socially assistive robot technology in elderly care". In: *BMJ open* 8.2 (2018), e018815 (cit. on pp. 10, 11).

[Ala+16] Alexandre Alahi et al. "Social lstm: Human trajectory prediction in crowded spaces". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 961–971 (cit. on pp. 47, 71).

[AS05] Julie A Adams and Marjorie Skubic. "Introduction to the special issue on human–robot interaction". In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 35.4 (2005), pp. 433–437 (cit. on p. 16).

[AS14] Henny Admoni and Brian Scassellati. "Nonverbal Behavior Modeling for Socially Assistive Robots". In: *2014 AAAI Fall Symposium Series*. 2014.

[Asa90] Haruhiko Asada. "Teaching and learning of compliance using neural nets: Representation and generation of nonlinear compliance". In: *Robotics and Automation, 1990. Proceedings., 1990 IEEE International Conference on*. IEEE. 1990, pp. 1237–1244 (cit. on p. 17).

[ASI01] Karen Allen, Barbara E Shykoff, and Joseph L Izzo. "Pet ownership, but not ACE inhibitor therapy, blunts home blood pressure responses to mental stress". In: *Hypertension* 38.4 (2001), pp. 815–820 (cit. on p. 10).

[Bac+11] Moez Baccouche et al. "Sequential deep learning for human action recognition". In: *International Workshop on Human Behavior Understanding*. Springer. 2011, pp. 29–39 (cit. on p. 103).

[Bad+02] Pierre Badin et al. "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images". In: *Journal of Phonetics* 30.3 (2002), pp. 533–553 (cit. on p. 89).

[Bai+06] Gérard Bailly et al. "Embodied conversational agents: computing and rendering realistic gaze patterns". In: *Pacific-Rim Conference on Multimedia*. Springer. 2006, pp. 9–18 (cit. on p. 89).

[Bai+16] Gérard Bailly et al. "Quantitative analysis of backchannels uttered by an interviewer during neuropsychological tests". In: *Interspeech*. 2016, pp. 2905–2909 (cit. on pp. 31, 33, 34, 61, 66, 68, 69, 85).

[Bar+07] Christoph Bartneck et al. "Is the uncanny valley an uncanny cliff?" In: *Robot and Human interactive Communication (RO-MAN), IEEE International Symposium on*. IEEE. 2007, pp. 368–373 (cit. on p. 98).

[BBA13] W Owen Brimijoin, Alan W Boyd, and Michael A Akeroyd. "The contribution of head movement to the externalization and internalization of sounds". In: *PloS one* 8.12 (2013), e83068 (cit. on p. 70).

[BER08] Gérard Bailly, Frédéric Elisei, and Stephan Raidt. "Boucles de perception-action et interaction face-à-face". In: *Revue française de linguistique appliquée* 13.2 (2008), pp. 121–131 (cit. on p. 37).

[Bet+16] Cindy L Bethel et al. "Using robots to interview children about bullying: Lessons learned from an exploratory study". In: *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE. 2016, pp. 712–717 (cit. on p. 14).

[Bev+07] Elisabetta Bevacqua et al. "An expressive ECA showing complex emotions". In: *Proceedings of the AISB annual convention, Newcastle, UK*. 2007, pp. 208–216 (cit. on p. 70).

[Bev+09] Frédéric Bevilacqua et al. "Continuous realtime gesture following and recognition". In: *International gesture workshop*. Springer. 2009, pp. 73–84 (cit. on p. 35).

[BF04] Christoph Bartneck and Jodi Forlizzi. "A design-centred framework for social human-robot interaction". In: *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*. IEEE. 2004, pp. 591–594 (cit. on p. 7).

[BG13] Aude Billard and Daniel Grollman. "Robot learning by demonstration". In: *Scholarpedia* 8.12 (2013), p. 3824 (cit. on p. 17).

[BH05] Gérard Bailly and Bleicke Holm. "SFC: a trainable prosodic model". In: *Speech communication* 46.3-4 (2005), pp. 348–364 (cit. on p. 85).

[Bha+17] Jaishankar Bharatharaj et al. "Robot-assisted therapy for learning and social interaction of children with autism spectrum disorder". In: *Robotics* 6.1 (2017), p. 4 (cit. on pp. 12, 14).

[Boe+02] Paulus Petrus Gerardus Boersma et al. "Praat, a system for doing phonetics by computer". In: *Glot international* 5 (2002), pp. 341–345 (cit. on pp. 28, 32).

[Bre04] Cynthia L Breazeal. *Designing sociable robots*. MIT press, 2004 (cit. on pp. 1, 17).

[BS02] Cynthia Breazeal and Brian Scassellati. "f 4 challenges in building robots that imitate people". In: *Imitation in animals and artifacts* 363 (2002) (cit. on p. 9).

[BS14] Raymond Brueckner and Bjorn Schulter. "Social signal classification using deep BLSTM recurrent neural networks". In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 4823–4827 (cit. on p. 46).

[BSS11] Cindy L Bethel, Matthew R Stevenson, and Brian Scassellati. "Secret-sharing: Interactions between a child, robot, and adult". In: *Systems, man, and cybernetics (SMC), 2011 IEEE International Conference on*. IEEE. 2011, pp. 2489–2494 (cit. on pp. 2, 10, 37).

[Bus+07] Carlos Busso et al. "Rigid head motion in expressive speech animation: Analysis and synthesis". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2007), pp. 1075–1086 (cit. on p. 70).

[Bus84] Herman Buschke. "Cued recall in amnesia". In: *Journal of Clinical and Experimental Neuropsychology* 6.4 (1984), pp. 433–440 (cit. on p. 23).

[BYSB13] Atef Ben Youssef, Hiroshi Shimodaira, and David A Braude. "Articulatory features for speech-driven head motion synthesis". In: *Proceedings of Interspeech, Lyon, France* (2013) (cit. on p. 70).

[CCD00] Alex Colburn, Michael F Cohen, and Steven Drucker. "The role of eye gaze in avatar mediated conversational interfaces". In: *Sketches and Applications, Siggraph'00* (2000) (cit. on p. 37).

[CCK03] Nicola Cathcart, Jean Carletta, and Ewan Klein. "A shallow model of backchannel continuers in spoken dialogue". In: *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics. 2003, pp. 51–58 (cit. on p. 60).

[CDP16] Lucie Cormons, Michel Dubois, and Caroline Poulet. "Compétences socio-communicatives mise en jeu entre un praticien et une personne âgée souffrant ou non de troubles neurocognitifs: quelles applications pour la robotique?" In: *Workshop sur les Affects, Compagnons Artificiels et Interactions (WACAI)*. Brest, France, 2016 (cit. on p. 98).

[CG+14] Álvaro Castro-González et al. "Learning Behaviors by an Autonomous Social Robot with Motivations". In: *Cybernetics and Systems* 45.7 (2014), pp. 568–598 (cit. on p. 17).

[CG02] Jacob W Crandall and Michael A Goodrich. "Characterizing efficiency of human robot interaction: A case study of shared-control teleoperation". In: *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*. Vol. 2. IEEE. 2002, pp. 1290–1295 (cit. on p. 21).

[CH92] Gregory F Cooper and Edward Herskovits. "A Bayesian method for the induction of probabilistic networks from data". In: *Machine learning* 9.4 (1992), pp. 309–347 (cit. on p. 52).

[Cho+15] François Chollet et al. *Keras*. 2015 (cit. on pp. 52, 63).

[Chu+17] Mei-Tai Chu et al. "Service innovation through social robot engagement to improve dementia care quality". In: *Assistive Technology* 29.1 (2017), pp. 8–18 (cit. on p. 11).

[Cil17] Carlo Ciliberto. "Connecting YARP to the Web with Yarp. js". In: *Frontiers in Robotics and AI* 4 (2017), p. 67 (cit. on p. 110).

[CVB01] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. "Beat: the behavior expression animation toolkit". In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM. 2001, pp. 477–486 (cit. on p. 40).

[Dau03] Kerstin Dautenhahn. "Roles and functions of robots in human society: implications from research in autism therapy". In: *Robotica* 21.4 (2003), pp. 443–452 (cit. on p. 6).

[Dea87] Thomas L Dean. "Intractability and time-dependent planning". In: *Reasoning About Actions & Plans*. Elsevier, 1987, pp. 245–266 (cit. on p. 41).

[Den] Britz Denny. *Recurrent Neural Networks Tutorial, Part 3 – Backpropagation Through Time and Vanishing Gradients*. Ed. by Motherboard. URL: `http://www.wildml.com/2015/10/recurrent-neural-networks-tutorial-part-3-backpropagation-through-time-and-vanishing-gradients/` (cit. on p. 46).

[DFC00] Catherine Dehon, Peter Filzmoser, and Christophe Croux. "Robust methods for canonical correlation analysis". In: *Data analysis, classification, and related methods*. Springer, 2000, pp. 321–326 (cit. on p. 72).

[Dio+15] Mélissa Dion et al. "Normative data for the Rappel libre/Rappel indicé à 16 items (16-item Free and Cued Recall) in the elderly Quebec-French population". In: *The Clinical Neuropsychologist* 28.sup1 (2015), pp. 1–19 (cit. on p. 30).

[DM12] M Dunham and K Murphy. "PMTK3: Probabilistic modeling toolkit for Matlab/Octave, version 3". In: *2012* (2012) (cit. on p. 51).

[DMBM15] Wim De Mulder, Steven Bethard, and Marie-Francine Moens. "A survey on the application of recurrent neural networks to statistical language modeling". In: *Computer Speech & Language* 30.1 (2015), pp. 61–98 (cit. on p. 46).

[Don+15] Jeffrey Donahue et al. "Long-term recurrent convolutional networks for visual recognition and description". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2625–2634 (cit. on p. 103).

[DPA13] Yu Ding, Catherine Pelachaud, and Thierry Artieres. "Modeling multimodal behaviors from speech prosody". In: *International Workshop on Intelligent Virtual Agents*. Springer. 2013, pp. 217–228 (cit. on p. 70).

[Fan+15] Yu Fang et al. "Eye-head coordination for visual cognitive processing". In: *PloS one* 10.3 (2015), e0121035 (cit. on p. 70).

[FBE15] François Foerster, Gérard Bailly, and Frédéric Elisei. "Impact of iris size and eyelids coupling on the estimation of the gaze direction of a robotic talking head by human viewers". In: *Humanoid Robots (Humanoids), IEEE-RAS 15th International Conference on*. IEEE. Portland, OR, 2015, pp. 148–153 (cit. on p. 89).

[FD06] Jodi Forlizzi and Carl DiSalvo. "Service robots in the domestic environment: a study of the roomba vacuum in the home". In: *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM. 2006, pp. 258–265 (cit. on p. 6).

[FKL14] Mary Ellen Foster, Simon Keizer, and Oliver Lemon. "Towards action selection under uncertainty for a socially aware robot bartender". In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM. 2014, pp. 158–159 (cit. on pp. 14, 15).

[FM13] Juan Fasola and Maja Mataric. "A socially assistive robot exercise coach for the elderly". In: *Journal of Human-Robot Interaction* 2.2 (2013), pp. 3–32 (cit. on p. 89).

[Fre01] Edward G Freedman. "Interactions between eye and head control signals can account for movement kinematics". In: *Biological cybernetics* 84.6 (2001), pp. 453–462 (cit. on p. 69).

[G.G+15] G.Gomez et al. "Qualitative assesment of a beaming environment for collaborative professional activities". In: (2015) (cit. on pp. 114, 115).

[GB18] Branislav Gerazov and Gérard Bailly. "PySFC-A System for Prosody Analysis based on the Superposition of Functional Contours Prosody Model". In: *9th International Conference on Speech Prosody (Speech Prosody 2018)*. ISCA. Poznan, Poland, 2018, pp. 774–778 (cit. on p. 85).

[GBX18] Branislav Gerazov, Gérard Bailly, and Yi Xu. "A Weighted Superposition of Functional Contours Model for Modelling Contextual Prominence of Elementary Prosodic Contours". In: *arXiv preprint arXiv:1806.06779* (2018) (cit. on p. 85).

[GHE04] Anders Green, Helge Huttenrauch, and K Severinson Eklundh. "Applying the Wizard-of-Oz framework to cooperative service discovery and configuration". In: *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*. IEEE. 2004, pp. 575–580 (cit. on p. 114).

[GJM13] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. "Hybrid speech recognition with deep bidirectional LSTM". In: *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE. 2013, pp. 273–278 (cit. on p. 46).

[GK02] Jennifer Goetz and Sara Kiesler. "Cooperation with a robotic assistant". In: *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2002, pp. 578–579 (cit. on p. 98).

[Gom+15] Guillermo Gomez et al. "Qualitative assessment of an immersive teleoperation environment for collaborative professional activities in a" beaming" experiment". In: *European conference for Virtual Reality and Augmented Reality (EuroVR)*. 2015, 8–pages (cit. on p. 115).

[Gra+02] Hans Peter Graf et al. "Visual prosody: Facial movements accompanying speech". In: *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE. 2002, pp. 396–401 (cit. on p. 70).

[Gra13] Alex Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013) (cit. on p. 128).

[GSS02] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. "Learning precise timing with LSTM recurrent networks". In: *Journal of machine learning research* 3.Aug (2002), pp. 115–143 (cit. on p. 47).

[Gui+15] G Guillermo et al. "Qualitative assesment of a beaming environment for collaborative professional activities". In: *European conference for Virtual Reality and Augmented Reality (EuroVR)*. 2015, p. 8 (cit. on p. 21).

[GV87] Daniel Guitton and Michel Volle. "Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range". In: *Journal of neurophysiology* 58.3 (1987), pp. 427–459 (cit. on p. 69).

[Hay+07] Kotaro Hayashi et al. "Humanoid robots as a passive-social medium-a field experiment at a train station". In: *Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on*. IEEE. 2007, pp. 137–144 (cit. on pp. 7, 8, 37).

[HC18] Ioannis Havoutis and Sylvain Calinon. "Learning from demonstration for semi-autonomous teleoperation". In: *Autonomous Robots* (2018), pp. 1–14 (cit. on p. 127).

[Heb+03] Liesi E Hebert et al. "Alzheimer disease in the US population: prevalence estimates using the 2000 census". In: *Archives of neurology* 60.8 (2003), pp. 1119–1122 (cit. on p. 5).

[Hee+09] Marcel Heerink et al. "Measuring acceptance of an assistive social robot: a suggested toolkit". In: *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*. IEEE. 2009, pp. 528–533 (cit. on p. 90).

[Hee+10] Marcel Heerink et al. "Assessing acceptance of assistive social agent technology by older adults: the almere model". In: *International journal of social robotics* 2.4 (2010), pp. 361–375 (cit. on p. 90).

[HK99] Martin Hansen and Birger Kollmeier. "Continuous assessment of time-varying speech quality". In: *The Journal of the Acoustical Society of America* 106.5 (1999), pp. 2888–2899 (cit. on p. 90).

[HLD15] Deanna Hood, Séverin Lemaignan, and Pierre Dillenbourg. "When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting". In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2015, pp. 83–90 (cit. on p. 87).

[HLY06] Salamin Hugues, Lmorency, and Song Yale. *HCRF library*. 2006 (cit. on p. 63).

[HM14] Chien-Ming Huang and Bilge Mutlu. "Learning-based modeling of multimodal behaviors for humanlike robots". In: *Human-robot interaction (HRI), ACM/IEEE international conference on*. ACM. 2014, pp. 57–64 (cit. on pp. 43, 44, 89).

[HMG10] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. "Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior". In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems. 2010, pp. 1265–1272 (cit. on p. 66).

[HNI] Daisuke Taniyama Hiroyuki Nitto and Hitomi Inagaki. *Social Acceptance and Impact of Robots and Artificial Intelligence*. Ed. by Nomura Research Institute (cit. on pp. 6–8).

[HR95] Roelof Hamberg and Huib de Ridder. "Continuous assessment of perceptual image quality". In: *Journal of the Optical Society of America* 12.12 (1995), pp. 2573–2577 (cit. on p. 90).

[HRJ04] Pamela J Hinds, Teresa L Roberts, and Hank Jones. "Whose job is it anyway? A study of human-robot interaction in a collaborative task". In: *Human-Computer Interaction* 19.1 (2004), pp. 151–181.

[HS16] Kathrin Haag and Hiroshi Shimodaira. "Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis". In: *International Conference on Intelligent Virtual Agents*. Springer. 2016, pp. 198–207 (cit. on p. 71).

[HS17] Mohamed K Helwa and Angela P Schoellig. "Multi-robot transfer learning: A dynamical system perspective". In: *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE. 2017, pp. 4702–4708 (cit. on p. 129).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 46).

[HU04] B. Hellwig and D. Uytvanck. "EUDICO Linguistic Annotator ELAN Version 2.0.2 manual". In: (2004) (cit. on p. 28).

[Ins03] Brent E Insko. "Measuring presence: Subjective, behavioral and physiological methods." In: (2003) (cit. on p. 118).

[JH16] Shafiq Joty and Enamul Hoque. "Speech act modeling of written asynchronous conversations with task-specific embeddings and conditional structured models". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2016, pp. 1746–1756 (cit. on p. 48).

[Jør+16] Nina Jøranson et al. "Change in quality of life in older people with dementia participating in Paro-activity: A cluster-randomized controlled trial". In: *Journal of advanced nursing* 72.12 (2016), pp. 3020–3033 (cit. on p. 11).

[Kan+02] Takayuki Kanda et al. "A constructive approach for developing interactive humanoid robots". In: *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*. Vol. 2. IEEE. 2002, pp. 1265–1270 (cit. on pp. 41, 42).

[Kan+03] Masao Kanamori et al. "Pilot study on improvement of quality of life among elderly using a pet-type robot". In: *Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on*. Vol. 1. IEEE. 2003, pp. 107–112 (cit. on p. 11).

[Kan+04] Takayuki Kanda et al. "Interactive robots as social partners and peer tutors for children: A field trial". In: *Human-computer interaction* 19.1 (2004), pp. 61–84 (cit. on p. 13).

[Kan+09] Takayuki Kanda et al. "An affective guide robot in a shopping mall". In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM. 2009, pp. 173–180 (cit. on pp. 8, 9, 15).

[KB07] Cory D Kidd and Cynthia Breazeal. "A robotic weight loss coach". In: *Proceedings of the national conference on artificial intelligence*. Vol. 22. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2007, p. 1985 (cit. on p. 10).

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 63).

[KCC10] Petar Kormushev, Sylvain Calinon, and Darwin G Caldwell. "Robot motor skill coordination with EM-based reinforcement learning". In: *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE. 2010, pp. 3232–3237 (cit. on p. 18).

[KFB09] Stephen C Kramer, Erika Friedmann, and Penny L Bernstein. "Comparison of the effect of human interaction, animal-assisted therapy, and AIBO-assisted therapy on long-term care residents with dementia". In: *Anthrozoös* 22.1 (2009), pp. 43–57 (cit. on p. 11).

[KFF15] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3128–3137 (cit. on p. 46).

[KH11] Iwan Kok and Dirk Heylen. "Observations on listener responses from multiple perspectives". In: *3rd Nordic Symposium on Multimodal Communication*. Northern European Association for Language Technology. Helsinki, Finland, 2011, pp. 48–55 (cit. on p. 90).

[KH12] Iwan de Kok and Dirk Heylen. "A survey on evaluation metrics for backchannel prediction models". In: *Feedback Behaviors in Dialog*. 2012 (cit. on p. 66).

[KI16] Takayuki Kanda and Hiroshi Ishiguro. *Human-robot interaction in social robotics*. CRC Press, 2016 (cit. on p. 13).

[Kie+08] Sara Kiesler et al. "Anthropomorphic interactions with a robot and robot–like agent". In: *Social Cognition* 26.2 (2008), pp. 169–181 (cit. on p. 7).

[Kim+13] Geon Ha Kim et al. "Structural brain changes after robot-assisted cognitive training in the elderly: A single-blind randomized controlled trial". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 9.4 (2013), P476–P477 (cit. on p. 11).

[KJ+15] Peter H Kahn Jr et al. "Will people keep the secret of a humanoid robot?: Psychological intimacy in hri". In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2015, pp. 173–180 (cit. on p. 10).

[Kop+06] Stefan Kopp et al. "Towards a common framework for multimodal generation: The behavior markup language". In: *International workshop on intelligent virtual agents*. Springer. 2006, pp. 205–217 (cit. on p. 40).

[KP04] Brigitte Krenn and Hannes Pirker. "Defining the gesticon: Language and gesture coordination for interacting embodied agents". In: *Proc. of the AISB-2004 Symposium on Language, Speech and Gesture for Expressive Characters*. 2004, pp. 107–115 (cit. on p. 81).

[KTT06] Cory D Kidd, Will Taggart, and Sherry Turkle. "A sociable robot to encourage social interaction among the elderly". In: *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE. 2006, pp. 3972–3976 (cit. on p. 11).

[Lem+13] Andre Lemme et al. "Kinesthetic teaching of visuomotor coordination for pointing by the humanoid robot icub". In: *Neurocomputing* 112 (2013), pp. 179–188 (cit. on p. 132).

[Lem+16] Séverin Lemaignan et al. "The Case of Handwriting". In: *IEEE Robotics & Automation Magazine* 1070.9932/16 (2016) (cit. on p. 10).

[LFS98] Shoudan Liang, Stefanie Fuhrman, and Roland Somogyi. "Reveal, a general reverse engineering algorithm for inference of genetic network architectures". In: (1998) (cit. on p. 52).

[LHZ17] Michal Luria, Guy Hoffman, and Oren Zuckerman. "Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. 2017, pp. 580–628 (cit. on p. 9).

[Li+15] Yanan Li et al. "Continuous role adaptation for human–robot shared control". In: *IEEE Transactions on Robotics* 31.3 (2015), pp. 672–681 (cit. on p. 127).

[Lin+04] Martial Van der Linden et al. "L'épreuve de rappel libre/rappel indicé à 16 items (RL/RI-16)". In: *L'évaluation des troubles de la mémoire : présentation de quatre tests de mémoire épisodique avec leur étalonnage*. Ed. by Martial Van der Linden. Marseille, France: Solal, 2004, pp. 25–47 (cit. on p. 23).

[Liu+12] Chaoran Liu et al. "Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction". In: *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*. IEEE. 2012, pp. 285–292 (cit. on p. 70).

[Liu+17] An-An Liu et al. "Benchmarking a multimodal and multiview and interactive dataset for human action recognition". In: *IEEE Transactions on cybernetics* 47.7 (2017), pp. 1781–1794 (cit. on p. 43).

[LM06] Jina Lee and Stacy Marsella. "Nonverbal behavior generator for embodied conversational agents". In: *International Workshop on Intelligent Virtual Agents*. Springer. 2006, pp. 243–255 (cit. on pp. 41, 70, 71).

[LMP01] John Lafferty, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: (2001) (cit. on p. 62).

[LTK09] Sergey Levine, Christian Theobalt, and Vladlen Koltun. "Real-time prosody-driven synthesis of body language". In: *ACM Transactions on Graphics (TOG)*. Vol. 28. 5. ACM. 2009, p. 172 (cit. on p. 70).

[Mac06] Beth L Macauley. "Animal-assisted therapy for persons with aphasia: A pilot study". In: *Journal of Rehabilitation Research and Development* 43.3 (2006), p. 357 (cit. on p. 11).

[Man+15] Jordan A Mann et al. "People respond better to robots than computer tablets delivering healthcare instructions". In: *Computers in Human Behavior* 43 (2015), pp. 112–117 (cit. on p. 10).

[Mas+10] Marcus Mast et al. "Semi-autonomous teleoperated learning in-home service robots for elderly care: A qualitative study on needs and perceptions of elderly people, family caregivers, and professional caregivers". In: *20th International Conference on Robotics and Mechatronics, Varna, Bulgaria, October 1-6*. 2010 (cit. on p. 128).

[Mas+15] Marcus Mast et al. "Design of the human-robot interaction for a semi-autonomous service robot to assist elderly people". In: *Ambient Assisted Living*. Springer, 2015, pp. 15–29 (cit. on pp. 127, 128).

[MB12] Soroosh Mariooryad and Carlos Busso. "Generating human-like behaviors using joint, speech-driven models for conversational agents". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.8 (2012), pp. 2329–2340 (cit. on p. 70).

[MBW15] A. Mihoub, G. Bailly, and C. Wolf. "Learning multimodal behavioral models for face-to-face social interaction". In: *Journal on Multimodal User Interfaces* 9.3 (2015), pp. 195–210 (cit. on pp. 43, 48, 49, 51–53, 58).

[Mel+09] Gail F Melson et al. "Children's behavior toward and understanding of robotic and living dogs". In: *Journal of Applied Developmental Psychology* 30.2 (2009), pp. 92–102 (cit. on p. 12).

[MFN06] Giorgio Metta, Paul Fitzpatrick, and Lorenzo Natale. "YARP: yet another robot platform". In: *International Journal of Advanced Robotic Systems* 3.1 (2006), p. 8 (cit. on p. 110).

[MHS17] Angelika Maier, Julian Hough, and David Schlangen. "Towards deep end-of-turn prediction for situated spoken dialogue systems". In: *Proceedings of INTERSPEECH 2017* (2017) (cit. on p. 61).

[Mih+16] Alaeddine Mihoub et al. "Graphical models for social behavior modeling in face-to face interaction". In: *Pattern Recognition Letters* 74 (2016), pp. 82–89 (cit. on pp. 25, 26, 43, 48, 51, 53, 54, 62, 71–73, 75).

[MKG10] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. "A probabilistic multimodal approach for predicting listener backchannels". In: *Autonomous Agents and Multi-Agent Systems* 20.1 (2010), pp. 70–84 (cit. on pp. 61, 62, 66).

[MMB15] Tobias May, Ning Ma, and Guy J Brown. "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 2679–2683 (cit. on p. 70).

[MN15] Yasser Mohammad and Toyoaki Nishida. *Data Mining for Social Robotics*. Springer, 2015 (cit. on pp. 16, 18).

[Mor10] Louis-Philippe Morency. "Modeling human communication dynamics [social sciences]". In: *IEEE Signal Processing Magazine* 27.5 (2010), pp. 112–116 (cit. on p. 60).

[Mor70] Masahiro Mori. "The uncanny valley". In: *Energy* 7.4 (1970), pp. 33–35 (cit. on pp. 89, 98).

[MPR08] Mehryar Mohri, Fernando Pereira, and Michael Riley. "Speech recognition with weighted finite-state transducers". In: *Springer Handbook of Speech Processing*. Springer, 2008, pp. 559–584 (cit. on p. 66).

[MR+18] Fernando Moya Rueda et al. "Convolutional Neural Networks for Human Activity Recognition Using Body-Worn Sensors". In: *Informatics*. Vol. 5. 2. Multidisciplinary Digital Publishing Institute. 2018, p. 26 (cit. on p. 103).

[Mub+13] Omar Mubin et al. "A review of the applicability of robots in education". In: *Journal of Technology in Education and Learning* 1.209-0015 (2013), p. 13 (cit. on p. 10).

[Mun+04] Kevin G Munhall et al. "Visual prosody and speech intelligibility: Head movement improves auditory speech perception". In: *Psychological science* 15.2 (2004), pp. 133–137 (cit. on p. 70).

[Mur+01] Kevin Murphy et al. "The bayes net toolbox for matlab". In: *Computing science and statistics* 33.2 (2001), pp. 1024–1034 (cit. on p. 52).

[NBE16] Duc-Canh Nguyen, Gérard Bailly, and Frédéric Elisei. "Conducting neuropsychological tests with a humanoid robot: design and evaluation". In: *Cognitive Infocommunications (CogInfoCom), IEEE International Conference on*. Warsaw, Poland, 2016, pp. 337–342 (cit. on pp. 68, 78, 92).

[NBE17] Duc-Canh Nguyen, Gérard Bailly, and Frédéric Elisei. "An Evaluation Framework to Assess and Correct the Multimodal Behavior of a Humanoid Robot in Human-Robot Interaction". In: *GEstures and SPeech in INteraction (GESPIN)*. 2017 (cit. on p. 68).

[NKN07] Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. "A spoken dialog system for chat-like conversations considering response timing". In: *Text, Speech and Dialogue*. Springer. 2007, pp. 599–606 (cit. on p. 60).

[Nod+14] Kuniaki Noda et al. "Multimodal integration learning of robot behavior using deep neural networks". In: *Robotics and Autonomous Systems* 62.6 (2014), pp. 721–736 (cit. on p. 43).

[Nom+05] Tatsuya Nomura et al. "Questionnaire–based research on opinions of visitors for communication robots at an exhibition in japan". In: *IFIP Conference on Human-Computer Interaction*. Springer. 2005, pp. 685–698 (cit. on p. 90).

[NS14] Ryoichi Nakashima and Satoshi Shioiri. "Why do we move our head to look at an object in our peripheral region? Lateral viewing interferes with attentive search". In: *PloS one* 9.3 (2014), e92284 (cit. on p. 70).

[OK09] Pierre-Yves Oudeyer and Frederic Kaplan. "What is intrinsic motivation? A typology of computational approaches". In: *Frontiers in neurorobotics* 1 (2009), p. 6 (cit. on p. 17).

[OKA06] Erhan Oztop, Mitsuo Kawato, and Michael Arbib. "Mirror neurons and imitation: A computationally guided review". In: *Neural Networks* 19.3 (2006), pp. 254–271 (cit. on p. 114).

[OR16] Francisco Javier Ordóñez and Daniel Roggen. "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition". In: *Sensors* 16.1 (2016), p. 115 (cit. on p. 47).

[OSY07] Kazuhiro Otsuka, Hiroshi Sawada, and Junji Yamato. "Automatic inference of cross-modal nonverbal interactions in multiparty conversations: who responds to whom, when, and how? from gaze, head gestures, and utterances". In: *Proceedings of the 9th international conference on Multimodal interfaces*. ACM. 2007, pp. 255–262 (cit. on p. 43).

[Par+15] Alberto Parmiggiani et al. "Design and Validation of a Talking Face for the iCub". In: *International Journal of Humanoid Robotics* 12.03 (2015), p. 1550026 (cit. on p. 85).

[Pas+11] Peter Pastor et al. "Skill learning and task outcome prediction for manipulation". In: *International Conference on Robotics and Automation (ICRA)*. IEEE. 2011, pp. 3828–3834 (cit. on pp. 17, 18).

[Pat+10] Ugo Pattacini et al. "An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots". In: *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE. 2010, pp. 1668–1674 (cit. on pp. 87, 131).

[Pol05] Martha E Pollack. "Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment". In: *AI magazine* 26.2 (2005), p. 9 (cit. on p. 5).

[PS08] Jan Peters and Stefan Schaal. "Reinforcement learning of motor skills with policy gradients". In: *Neural networks* 21.4 (2008), pp. 682–697 (cit. on p. 18).

[PY+10] Sinno Jialin Pan, Qiang Yang, et al. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2010), pp. 1345–1359 (cit. on p. 128).

[Qi+17] Siyuan Qi et al. "Predicting human activities using stochastic grammar". In: *International Conference on Computer Vision (ICCV), IEEE*. 2017 (cit. on p. 103).

[Qur+16] Ahmed Hussain Qureshi et al. "Robot gains social intelligence through multimodal deep reinforcement learning". In: *Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on*. IEEE. 2016, pp. 745–751 (cit. on p. 18).

[Rah+16] Rouhollah Rahmatizadeh et al. "From virtual demonstration to real-world manipulation using LSTM and MDN". In: *arXiv preprint arXiv:1603.03833* (2016) (cit. on p. 18).

[Rav+16] Harish Chaandar Ravichandar et al. "Learning and predicting sequential tasks using recurrent neural networks and multiple model filtering". In: *2016 AAAI Fall Symposium Series*. 2016 (cit. on p. 48).

[RDS08] Daniel Richardson, Rick Dale, and Kevin Shockley. "Synchrony and swing in conversation: Coordination, temporal dynamics, and communication". In: *Embodied communication in humans and machines* (2008), pp. 75–94 (cit. on p. 58).

[RMB14] Hayley Robinson, Bruce MacDonald, and Elizabeth Broadbent. "The role of healthcare robots for older people at home: a review". In: *International Journal of Social Robotics* 6.4 (2014), pp. 575–591 (cit. on p. 9).

[RMB15] Hayley Robinson, Bruce MacDonald, and Elizabeth Broadbent. "Physiological effects of a companion robot on blood pressure of older people in residential care facility: A pilot study". In: *Australasian journal on ageing* 34.1 (2015), pp. 27–32 (cit. on p. 11).

[RMK14] Pranav Rane, Varun Mhatre, and Lakshmi Kurup. "Study of a home robot: Jibo". In: *International journal of engineering research and technology* 3.10 (2014) (cit. on p. 7).

[RN96] Byron Reeves and Clifford Ivar Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press, 1996 (cit. on p. 5).

[Rob+13] Hayley Robinson et al. "The psychosocial effects of a companion robot: a randomized controlled trial". In: *Journal of the American Medical Directors Association* 14.9 (2013), pp. 661–667 (cit. on p. 11).

[Rob16] Blue Frog Robotics. "Buddy the first companion robot". In: *Retrieved* 8 (2016), p. 2016 (cit. on p. 7).

[Ron+16] Alessandro Roncone et al. "A Cartesian 6-DoF Gaze Controller for Humanoid Robots." In: *Robotics: Science and Systems*. Vol. 2016. 2016 (cit. on pp. 88, 108, 131).

[Ros14] Benjamin Rosman. "Behavioural domain knowledge transfer for autonomous agents". In: *2014 AAAI Fall Symposium Series*. 2014 (cit. on p. 128).

[Roy+00] Nicholas Roy et al. "Towards personal service robots for the elderly". In: *Workshop on Interactive Robots and Entertainment (WIRE 2000)*. Vol. 25. 2000, p. 184 (cit. on p. 5).

[Rue+17] Robin Ruede et al. "Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor". In: *arXiv preprint arXiv:1706.01340* (2017) (cit. on p. 61).

[Ryb+07] Paul E Rybski et al. "Interactive robot task training through dialog and demonstration". In: *2nd International Conference on Human-Robot Interaction (HRI)*. IEEE. 2007, pp. 49–56 (cit. on p. 18).

[Sab+13] Selma Sabanovic et al. "PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia". In: *Rehabilitation Robotics (ICORR), 2013 IEEE International Conference on*. IEEE. 2013, pp. 1–6 (cit. on p. 11).

[Sae+10] Martin Saerbeck et al. "Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2010, pp. 1613–1622 (cit. on p. 10).

[Sai+03] Tomoko Saito et al. "Relationship between interaction with the mental commit robot and change of stress reaction of the elderly". In: *Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on*. Vol. 1. IEEE. 2003, pp. 119–124 (cit. on p. 11).

[San+07] David Sander et al. "Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion". In: *European Journal of Cognitive Psychology* 19.3 (2007), pp. 470–480 (cit. on p. 89).

[SB03] Jean Scholtz and Siavosh Bahrami. "Human-robot interaction: development of an evaluation methodology for the bystander role of interaction". In: *Systems, Man and Cybernetics, 2003. IEEE International Conference on*. Vol. 4. IEEE, 2003, pp. 3212–3217 (cit. on p. 89).

[SB17] Najmeh Sadoughi and Carlos Busso. "Speech-driven animation with meaningful behaviors". In: *arXiv preprint arXiv:1708.01640* (2017) (cit. on p. 71).

[Sch+18] Paul Schydlo et al. "Anticipation in Human-Robot Cooperation: A recurrent neural network approach for multiple action sequences prediction". In: *arXiv preprint arXiv:1802.10503* (2018) (cit. on p. 48).

[Sch06] Erin E Schultz. "Furry Therapists: The Advantages and Disadvantages of Implementing". In: *Social Work* 60 (2006), p. 64 (cit. on p. 10).

[Seq+16] Pedro Sequeira et al. "Discovering social interaction strategies for robots from restricted-perception Wizard-of-Oz studies". In: *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*. IEEE. 2016, pp. 197–204 (cit. on p. 138).

[Shi+06] Masahiro Shiomi et al. "Interactive humanoid robots for a science museum". In: *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM. 2006, pp. 305–312 (cit. on pp. 8, 10, 14).

[Shi+15] Kyriacos Shiarlis et al. "TERESA: a socially intelligent semi-autonomous telepresence system". In: *Workshop on machine learning for social robotics at ICRA-2015 in Seattle*. 2015 (cit. on p. 128).

[SJS09] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. "The oz of wizard: simulating the human for interaction research". In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM. 2009, pp. 101–108 (cit. on p. 105).

[Ska17] Gabriel Skantze. "Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 2017, pp. 220–230 (cit. on p. 61).

[SKH11] Alessandra Maria Sabelli, Takayuki Kanda, and Norihiro Hagita. "A conversational robot in an elderly care center: an ethnographic study". In: *Proceedings of the 6th international conference on Human-robot interaction*. ACM. 2011, pp. 37–44 (cit. on pp. 12, 13).

[SM+12] Charles Sutton, Andrew McCallum, et al. "An introduction to conditional random fields". In: *Foundations and Trends® in Machine Learning* 4.4 (2012), pp. 267–373 (cit. on p. 62).

[Sof+14] B Sofiane et al. "Learning of social signatures through imitation game between a robot and a human partner, Autonomous Mental Development". In: *IEEE Transactions on* 6.3 (2014), pp. 213–225 (cit. on p. 18).

[SP03] Fei Sha and Fernando Pereira. "Shallow parsing with conditional random fields". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics. 2003, pp. 134–141 (cit. on p. 62).

[Spe+14] Patric R Spence et al. "Welcoming our robot overlords: Initial expectations about interaction with a robot". In: *Communication Research Reports* 31.3 (2014), pp. 272–280.

[Ste+12] Anthony Steed et al. "Beaming: an asymmetric telepresence system". In: *IEEE computer graphics and applications* 32.6 (2012), pp. 10–17 (cit. on p. 115).

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. 2014, pp. 3104–3112 (cit. on p. 46).

[Tan+06] Fumihide Tanaka et al. "Daily HRI evaluation at a classroom environment: reports from dance interaction experiments". In: *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 2006, pp. 3–9 (cit. on p. 90).

[Tan+12] Masaaki Tanaka et al. "Effect of a human-type communication robot on cognitive function in elderly women living alone". In: *Medical science monitor: international medical journal of experimental and clinical research* 18.9 (2012), CR550 (cit. on p. 11).

[Tap09] Adriana Tapus. "Improving the quality of life of people with dementia through the use of socially assistive robots". In: *Advanced Technologies for Enhanced Quality of Life, 2009. AT-EQUAL'09.* IEEE. 2009, pp. 81–86 (cit. on p. 11).

[TBW16] Eleni Tsironi, Pablo Barros, and Stefan Wermter. "Gesture recognition with a convolutional long short-term memory recurrent neural network". In: *Bruges, Belgium* 2 (2016) (cit. on p. 47).

[Thó02] Kristinn R Thórisson. "Natural turn-taking needs no manual: Computational theory and model, from perception to action". In: *Multimodality in language and speech systems*. Springer, 2002, pp. 173–207 (cit. on p. 41).

[Thó99] Kristinn R Thórisson. "Mind model for multimodal communicative creatures and humanoids". In: *Applied Artificial Intelligence* 13.4-5 (1999), pp. 449–486 (cit. on pp. 37, 38).

[TM08] Adriana Tapus and Maja J Mataric. "Socially Assistive Robots: The Link between Personality, Empathy, Physiological Signals, and Task Performance." In: *AAAI spring symposium: emotion, personality, and social behavior*. 2008, pp. 133–140 (cit. on p. 112).

[TML15] Leimin Tian, Johanna D Moore, and Catherine Lai. "Emotion recognition in spontaneous and acted dialogues". In: *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE. 2015, pp. 698–704 (cit. on p. 47).

[TPH10] Khiet P Truong, RW Poppe, and DKJ Heylen. "A rule-based backchannel prediction model using pitch and pause information". In: (2010) (cit. on p. 60).

[VDB+10] Jur Van Den Berg et al. "Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations". In: *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE. 2010, pp. 2074–2081.

[VNK15] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. "A review of human activity recognition methods". In: *Frontiers in Robotics and AI* 2 (2015), p. 28 (cit. on p. 43).

[Vog+14] David Vogt et al. "Learning two-person interaction models for responsive synthetic humanoids". In: *Journal of Virtual Reality and Broadcasting* 11.1 (2014) (cit. on p. 43).

[Vol+16] Anna-Lisa Vollmer et al. "Pragmatic frames for teaching and learning in human–robot interaction: Review and challenges". In: *Frontiers in neurorobotics* 10 (2016), p. 10 (cit. on p. 129).

[VS+15] Meritxell Valentí Soler et al. "Social robots in advanced dementia". In: *Frontiers in aging neuroscience* 7 (2015), p. 133 (cit. on p. 11).

[Wad+03] Kazuyoshi Wada et al. "Relationship between familiarity with mental commit robot and psychological effects to elderly people by robot assisted activity". In: *Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on*. Vol. 1. IEEE. 2003, pp. 113–118 (cit. on p. 11).

[Wad+08] Kazuyoshi Wada et al. "Robot therapy for elders affected by dementia". In: *IEEE Engineering in medicine and biology magazine* 27.4 (2008) (cit. on p. 11).

[Wal+08] Michael L Walters et al. "Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion". In: *Autonomous Robots* 24.2 (2008), pp. 159–178.

[Wan+18] Yuxuan Wang et al. "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis". In: *arXiv preprint arXiv:1803.09017* (2018) (cit. on pp. 35, 128).

[Wer90] Paul J Werbos. "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10 (1990), pp. 1550–1560 (cit. on p. 45).

[Wit+06] Peter Wittenburg et al. "Elan: a professional framework for multimodality research". In: *international conference on Language Resources and Evaluation (LREC)*. 2006, pp. 1556–1559 (cit. on p. 32).

[WS07] Kazuyoshi Wada and Takanori Shibata. "Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house". In: *IEEE Transactions on Robotics* 23.5 (2007), pp. 972–980 (cit. on p. 12).

[WT00] Nigel Ward and Wataru Tsukahara. "Prosodic features which cue back-channel responses in English and Japanese". In: *Journal of pragmatics* 32.8 (2000), pp. 1177–1207 (cit. on p. 60).

[Xin+15] SHI Xingjian et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting". In: *Advances in neural information processing systems*. 2015, pp. 802–810 (cit. on p. 103).

[YB07] Li Yujian and Liu Bo. "A normalized Levenshtein distance metric". In: *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), pp. 1091–1095 (cit. on p. 53).

[YD02] Holly A Yanco and Jill L Drury. "A taxonomy for human-robot interaction". In: *Proceedings of the AAAI Fall Symposium on Human-Robot Interaction*. 2002, pp. 111–119 (cit. on p. 21).

[YKVB00] Hani Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. "Facial animation and head motion driven by speech acoustics". In: *5th Seminar on Speech Production: Models and Data*. Kloster Seeon, Germany. 2000, pp. 265–268 (cit. on p. 70).

[You+11] James E Young et al. "Evaluating human-robot interaction". In: *International Journal of Social Robotics* 3.1 (2011), pp. 53–67 (cit. on p. 89).

[ZWM15] Minhua Zheng, Jiaole Wang, and Max Q-H Meng. "Comparing two gesture design methods for a humanoid robot: Human motion mapping by an RGB-D sensor and hand-puppeteering". In: *Robot and Human Interactive Communication (RO-MAN), 24th IEEE International Symposium on*. IEEE. 2015, pp. 609–614 (cit. on p. 89).

[ZZ14] Min-Ling Zhang and Zhi-Hua Zhou. "A review on multi-label learning algorithms". In: *IEEE transactions on knowledge and data engineering* 26.8 (2014), pp. 1819–1837 (cit. on p. 52).

Resumé/Abstract

Résumé — Un robot d'assistance sociale (SAR) est destiné à engager les gens dans une interaction située comme la surveillance de l'exercice physique, la réadaptation neuropsychologique ou l'entraînement cognitif. Alors que les comportements interactifs de ces systèmes sont généralement scriptés, nous discutons ici du cadre d'apprentissage de comportements interactifs multimodaux qui est proposé par le projet SOMBRERO. Dans notre travail, nous avons utilisé l'apprentissage par démonstration afin de fournir au robot des compétences nécessaires pour effectuer des tâches collaboratives avec des partenaires humains. Il y a trois étapes principales d'apprentissage de l'interaction par démonstration: (1) recueillir des comportements interactifs représentatifs démontrés par des tuteurs humains; (2) construire des modèles des comportements observés tout en tenant compte des connaissances a priori (modèle de tâche et d'utilisateur, etc.); et ensuite (3) fournir au robot-cible des contrôleurs de gestes appropriés pour exécuter les comportements souhaités. Les modèles multimodaux HRI (Human-Robot Interaction) sont fortement inspirés des interactions humain-humain (HHI). Le transfert des comportements HHI aux modèles HRI se heurte à plusieurs problèmes: (1) adapter les comportements humains aux capacités interactives du robot en ce qui concerne ses limitations physiques et ses capacités de perception, d'action et de raisonnement limitées; (2) les changements drastiques des comportements des partenaires humains face aux robots ou aux agents virtuels; (3) la modélisation des comportements interactifs conjoints; (4) la validation des comportements robotiques par les partenaires humains jusqu'à ce qu'ils soient perçus comme adéquats et significatifs. Dans cette thèse, nous étudions et faisons des progrès sur ces quatre défis. En particulier, nous traitons les deux premiers problèmes (transfert de HHI vers HRI) en adaptant le scénario et en utilisant la téléopération immersive. En outre, nous utilisons des réseaux neuronaux récurrents pour modéliser les comportements interactifs multimodaux (tels que le discours, le regard, les mouvements de bras, les mouvements de la tête, les canaux). Ces techniques récentes surpassent les méthodes traditionnelles (Hidden Markov Model, Dynamic Bayesian Network, etc.) en termes de précision et de coordination inter-modalités. A la fin de cette thèse, nous évaluons une première version de robot autonome équipé des modèles construits par apprentissage.

Mots clés : Socially Assistive Robot, Comportements Interactifs Multimodaux, LSTM, Téléopération Immersive, Apprendre par Démonstration

Abstract — A socially assistive robot (SAR) is meant to engage people into situated interaction such as monitoring physical exercise, neuropsychological rehabilitation or cognitive training. While the interactive behavioral policies of such systems are mainly hand-scripted,

we discuss here key features of the training of multimodal interactive behaviors in the framework of the SOMBRERO project. In our work, we used learning by demonstration in order to provide the robot with adequate skills for performing collaborative tasks in human centered environments. There are three main steps of learning interaction by demonstration: we should (1) collect representative interactive behaviors from human coaches; (2) build comprehensive models of these overt behaviors while taking into account a priori knowledge (task and user model, etc.); and then (3) provide the target robot with appropriate gesture controllers to execute the desired behaviors. Multimodal HRI (Human-Robot Interaction) models are mostly inspired by Human-Human interaction (HHI) behaviors. Transferring HHI behaviors to HRI models faces several issues: (1) adapting the human behaviors to the robot's interactive capabilities with regards to its physical limitations and impoverished perception, action and reasoning capabilities; (2) the drastic changes of human partner behaviors in front of robots or virtual agents; (3) the modeling of joint interactive behaviors; (4) the validation of the robotic behaviors by human partners until they are perceived as adequate and meaningful. In this thesis, we study and make progress over those four challenges. In particular, we solve the two first issues (transfer from HHI to HRI) by adapting the scenario and using immersive teleoperation. In addition, we use Recurrent Neural Networks to model multimodal interactive behaviors (such as speech, gaze, arm movements, head motion, backchannels) that surpass traditional methods (Hidden Markov Model, Dynamic Bayesian Network, etc.) in both accuracy and coordination between the modalities. We also build and evaluate a proof-of-concept autonomous robot to perform the tasks.

Keywords: Socially Assistive Robot, Multimodal Interactive Behaviors, LSTM, Immersive Teleoperation, Learning by Demonstration

---

GIPSA-lab, 11 rue des Mathématiques, 38402 Saint-Martin-d'Heres CEDEX