



Konputazio-Zientziak eta Adimen Artifiziala Saila  
Informatika Fakultatea

# Towards a Framework for Socially Interactive Robots

**PhD Thesis**

Igor Rodriguez Rodriguez

**Supervision**

Elena Lazkano Ortega  
Txelo Ruiz Vazquez

Dissertation submitted to the Department of Computer Science and Artificial  
Intelligence of the University of the Basque Country (UPV/EHU) as partial fulfilment  
of the requirements for the PhD degree in Computer Science

October 2018





Bittor Camara Hierro, *“The Sleep of Reason Produces Robot Monsters”*.

Aurrerapena lagun  
datorren garaia,  
gizaki ta makinez  
doa bidaia,  
munduko aldaketek  
marratzuz paisaia,  
teknologiaren dantzan  
geroen usaia,  
badator anaia,  
zientzien talaia,  
roboten kuraia,  
ikastea nahia,  
iraultzari heltzeko  
gaitezen saia.

Mikel Esteban Urriaga





# Acknowledgements

Legend goes that carrying out a doctoral dissertation is like a roller coaster, a journey full of ups and downs from which you can not get off until the end. Now I can confirm that such legend is true. It is beautiful, crazy, exciting, frustrating, amazing and difficult all at the same time. It has been a long journey to get here, but finally, the beginning has come to an end. Although honestly, I doubt that this is a real end.

First of all I would like to give special thanks to “The Brikos” team, my supervisors Elena Lazkano and Txelo Ruiz, for the chance they gave me to carry out this dissertation. You both gave me the opportunity to be introduced into the world of robots, and to learn and acquire the knowledge that I have today about these brilliant machines. Thanks Elena for transmitting your passion for robots to me, and Txelo for trying to teach me your knowledge as expert robot technician.

Of course, this adventure would not have been possible without the RSAIT team colleagues. I would like to thank: Basi (The Master Boss) for the opportunity he gave me when he accepted me in the group, Iñigo (Indigo) and Ekaitz (Bilbaino) for their valuable advice, and specially Ozteta and Aitzol, for supporting and helping me during the last phase of the research. I owe you a couple of beers!

Neither can I forget about my battle buddies Oihane, Florian, and Adriano. Their help during the development of some parts of this work as well as the funny moments we share together are very valuable for me. Thanks to CRSS lab people too for their hospitality during my stay in Palermo.

I would also like to thank my friends for supporting me in good and hard times and making me forget about work when I was stressed. *Eskerrik asko Pinupe! pintxopoterapiak asko lagundu dozta azkeneko txanpa honetan. Mila esker zuei ere Tinkiwinkis! Eta batez ere, esker anitz zuei Bittor eta Mikel!! Zuek bai artista galantak.*

Finally, and most importantly, I would like to thank my family. While developing this work I had good moments but also difficult periods in which they have always been close. My parents deserve a special mention: *Ama y Aita, eskerrik asko por vuestro afecto, apoyo y comprensión, pero sobre todo gracias por todo el esfuerzo que habéis hecho para que pueda cumplir mis sueños.*

Once again, thank you all for joining me during this journey on the roller coaster of my thesis.



# Abstract

Over the most recent decades, research in the field of social robotics has considerably grown up. There are a growing number of different types of robots, and their roles within society are expanding little by little. Robots endowed with social abilities aim to be used for different applications; for instance, as interactive teachers and educational assistants, to support diabetes management in children, to assist elderly with special needs, as interactive characters in theatre, or even as hotel and shopping mall assistants.

The RSAIT research team has worked on several areas of robotics, particularly, in control architectures, robot exploration and navigation, machine learning, and computer vision. This new research work aims to add a new layer to the previous development, the layer of human-robot interaction that focuses on the social capabilities that a robot must show while interacting with people, such as express and perceive emotions, communicate with high-level dialogue, learn models of other agents, establish and maintain social relationships, use natural cues (gaze, gestures, etc.), show distinctive personality and character and learn social competencies.

In this dissertation, we have tried to bear our grain of sand to the basic questions that arise when thinking about social robots: (1) How do we, humans, communicate with (or operate) social robots?; and (2) How do social robots act with us? In that vein, the work has been developed in two phases: in the first phase we have focused on exploring from a practical point of view several ways that humans use to communicate with robots in a natural manner. Additionally, in the second phase, we have investigated on how social robots must act with the user.

With respect to the first phase, three natural user interfaces intended to make the interaction with social robots more natural have been developed. Those interfaces have been tested by developing two applications of different use: guide robots and a humanoid robot control system for entertainment. Working on those applications allowed us to endow our robots with some basic skills, such as navigation, inter-robot communication and speech recognition and understanding capabilities.

On the other hand, in the second phase we have focused on identifying and developing the basic behavioural modules that this type of robots need to be socially believable and trustworthy while acting as social agents. We presented a framework for socially interactive robots that allows the robot to express (kind of) emotions and show a natural human-like body language according to the task to be performed and the environmental conditions.

The validation of the different states of development of our social robots is done

in public representations. Exposing our robots to the public in those performances has become an essential tool for qualitatively measuring the social acceptance of the prototypes being developed. In the same way robots need a physical body to interact with the environment and to become intelligent, social robots need to socially participate in the real tasks they are being built for in order to improve their sociability.

# Contents

<b>I</b>	<b>Introduction</b>	<b>1</b>
<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Goals . . . . .	4
1.2	Robotic platforms . . . . .	6
1.3	Structure of the document . . . . .	9
<b>II</b>	<b>First steps for social interaction with robots</b>	<b>11</b>
<b>2</b>	<b>Interacting with tour guide robots</b>	<b>13</b>
2.1	Service robots . . . . .	14
2.2	Tour guide robots . . . . .	15
2.3	Robot navigation . . . . .	16
2.4	Contributions and conclusions . . . . .	18
<b>3</b>	<b>Interacting with a humanoid robot through Natural User Interfaces</b>	<b>23</b>
3.1	Remotely controlled robots . . . . .	23
3.2	Natural user interfaces for remote controlled robots . . . . .	25
3.3	Contributions and conclusions . . . . .	29
<b>III</b>	<b>A framework for socially interactive robots</b>	<b>33</b>
<b>4</b>	<b>Socially interactive robots</b>	<b>35</b>
4.1	Robots on stage . . . . .	36
4.2	Bertsolaritza . . . . .	36
4.3	The ecological niche of Bertsobot . . . . .	38
<b>5</b>	<b>Describing the basic behaviours of Bertsobot</b>	<b>39</b>
5.1	Verbal communication . . . . .	40
5.2	Perception of the environment . . . . .	43
5.3	Non-Verbal communication: Body language . . . . .	45
5.4	Contributions and conclusions . . . . .	48
<b>6</b>	<b>Improving autonomy of social robots by self-awareness</b>	<b>51</b>

6.1	Self-body awareness . . . . .	52
6.2	Contributions and conclusions . . . . .	55
<b>7</b>	<b>Making our robots affective</b>	<b>59</b>
7.1	Emotion theories . . . . .	60
7.2	Understanding and reacting to audience feedback . . . . .	61
7.3	Talking with sentiment . . . . .	65
7.4	Contributions and conclusions . . . . .	70
<b>8</b>	<b>Enhancing spontaneity</b>	<b>73</b>
8.1	Generative models and their applications . . . . .	73
8.2	Beyond deterministic body language . . . . .	75
8.3	Combining the gesture generation system with the Adaptive Talking Behaviour . . . . .	78
8.4	Contributions and conclusions . . . . .	80
<b>9</b>	<b>A framework for socially interacting robots</b>	<b>81</b>
9.1	Stages of the architecture . . . . .	82
9.2	Evolution of the system through public performances . . . . .	84
<b>IV</b>	<b>Conclusions and further work</b>	<b>89</b>
<b>10</b>	<b>Conclusions and further work</b>	<b>91</b>
10.1	Conclusions . . . . .	91
10.2	Autonomy of our social robots . . . . .	93
10.3	Further work . . . . .	95
<b>V</b>	<b>Publications</b>	<b>97</b>
<b>11</b>	<b>Publications related to Part II</b>	<b>99</b>
11.1	Standardization of a Heterogeneous Robots Society Based on ROS . . .	99
11.2	GidaBot: a system of heterogeneous robots collaborating as guides in multi-floor environments . . . . .	125
11.3	Humanizing NAO robot teleoperation using ROS . . . . .	137
11.4	NAO Robot as Rehabilitation Assistant in a Kinect Controlled System	146
<b>12</b>	<b>Publications related to Part III</b>	<b>152</b>
12.1	Singing minstrel robots, a means for improving social behaviors . . .	152
12.2	Minstrel robots: Body language expression through applause evaluation . . . . .	159
12.3	BertsoBot: Towards a Framework for Socially Interacting Robots . .	166
12.4	Body Self-Awareness for Social Robots . . . . .	173
12.5	On how self-body awareness improves autonomy in social robots . . .	179

12.6 Adaptive Emotional Chatting Behavior to Increase the Sociability of Robots . . . . .	186
12.7 Spontaneous Talking Gestures Using Generative Adversarial Networks	197
12.8 Talking with Sentiment: Adaptive Expression Generation Behavior for Social Robots . . . . .	219
12.9 Robots on Stage: a Cognitive Framework for Socially Interacting Robots	226





Part I  
Introduction



# Chapter 1

## Introduction

On April 18, 2012, the dean of the University of the Basque Country organised a unique event of Basque improvised poetry (called *bertsolaritza*), a performance of *bertsolari-s* and robots. Galtxagorri and Tartalo – two standard wheeled platforms – shared the stage with several professional verse-makers or *bertsolari-s*, and performed in front of an audience. The performance aroused great interest, and almost every local newspaper, radio and television covered the event (see Figure 1.1 and videos<sup>1,2</sup>). Several researchers from the Faculty of Informatics worked together to meet the challenge, and the work developed for the event turned out to be the first prototype of a *bertsolari* robot [7].



Figure 1.1: Speaker’s corner inauguration event

The event, which was initially meant as a “game”, placed us in a completely new scenario for the robots: the robots we had used so far for navigational tasks were placed in a new environment and had to perform a task in which it was essential to interact with people. Several lessons were extracted from that event. The first one

---

<sup>1</sup><https://www.youtube.com/watch?v=x8w4YuNY-Z0>

<sup>2</sup><https://www.youtube.com/watch?v=OpQBVmkzRWg&t=82s>

was related to the robot morphology: despite the empathy that the public showed towards the robots, we realised that the robots used were not entirely suitable for social interaction due to their small number of DoF. Their expressiveness was limited to small oscillations and camera movements simulating dancing and head movements while singing. The second lesson learnt involved verbal communication: even though the improvised poems proved to be an effective communication tool, methods were needed to enhance verse coherence, as well as to make the robots understand verbal orders and give appropriate responses. Finally, the autonomy level of the robots on the stage had to be increased and, more importantly, the way the robots behave on the stage had to be humanised.

What started as a challenge for a special event, became a new line of investigation for the RSAIT research group, with a practical purpose: the design and development of a *bertsolari* robot, which we called Bertsobot. But whereas the main task of a troubadour robot is to process the verbal instructions of the presenter and to compose the best possible verse, the main reason that motivated this research work, was not to create a machine able to compose verses and sing them but to develop a robot with social skills able to understand verbal instructions, navigate around, recognise the key elements of its surroundings and interact in a natural way with other agents and the audience, showing the appropriate degree of expressiveness.

We believe that such stage performance is valuable both as an implementation platform and as a testing ground for Social Robotics (SR) research. On the one hand, the performance setting is constrained to some degree as it limits the perception and actuation possibilities of the robotic system. On the other hand, it provides a unique environment in which humans and robots collaborate and where dialog, sensory processing, action selection and behaviour coordination are required.

This dissertation proposes a framework for Socially Interactive Robots (SIR) [59] composed by several robot behaviours that endow the robots with some Human-Robot Interaction (HRI) abilities needed in social interactions: expressing and perceiving emotion, and communicating at high-level dialogue and using natural cues (gaze, gestures, etc.).

## 1.1 Goals

The RSAIT research team has worked in several areas of robotics, particularly, in control architectures, robot exploration and navigation, machine learning, and computer vision. Safe navigation in indoor environments is one of the group's main goal [97], avoiding obstacles, identifying objectives, exploring new places [83], etc.

This new research work aims to add a new layer to the previous work, the layer of HRI that focuses on social interaction. This means that we need to take the robots out of the laboratory, place them in complex human contexts, and prepare them to understand human orders and expressions, as well as to act appropriately showing human-like behaviours.

The Bertsobot project provided us with the opportunity to get an in-depth view

of the area of social robotics, and brought about new challenges. Taking the robots out of the faculty and putting them at the service of people marked a milestone for the RSAIT<sup>3</sup> group.

Given that the word *interaction* implies a two-way action, this research work intends to contribute to two basic questions that arise when thinking about social robots:

1. How do we, humans, communicate with (or operate) social robots?
2. How do social robots act with us?

Our first steps will focus on a more classic HRI research, which aims to design interfaces that enable a person to process and understand the state of the robot and to simultaneously provide the robot with appropriate movement and action commands. We will start experimenting with three different ways to communicate with robots:

- Through Graphical User Interfaces (GUIs).
- By using the human operator's body as a controller.
- By voice commands.

This experimentation is done by developing two robot applications for different purposes: guide robots, that can be categorised as service robots with some social capabilities; and a Natural User Interface (NUI) for entertainment.

In addition to the interfaces for interaction developed, working on the two applications should allow us to endow our robots with some basic skills, such as navigation, inter-robot communication, and speech recognition and understanding capabilities. This phase should also help us gain knowledge and a better understanding of the capabilities the robots can show.

Afterwards, we shall move on to explore the idea of social and affective interaction with robots, focusing on the behaviours that they must show when interacting with the user. The main purpose of this step is to develop the basic behavioural modules that robots need to be socially believable and trustworthy when they act as social agents. We would like to take the communication and interaction between humans and robots one step further by developing a framework (architecture) that will allow robots to:

- Have a conversation with other agents, understanding their requirements and providing appropriate responses.
- Have a perception of the environment, identifying and recognising objects and other agents in the environment.
- Have a perception of their own body configuration.

---

<sup>3</sup><http://www.sc.ehu.es/ccwrobot/seccion/home/lang/en>

- Obtain emotional feedback from the environment.
- Express (some kind of) emotions and show a natural human-like body language according to their emotional state, the task to be performed and the environmental conditions.

Here, we propose *bertsolaritza* as test-bed application. Robots acting as human troubadours in public events is an ambitious aim that entails a complex scenario. In our case, the behaviours to be developed are limited to some degree by the *bertsolaritza* context, the task that the verse-maker has to perform (speak, sing, react, etc.) and the sensory-motor capabilities of the robots.

The work proposed has also an extra goal: to disseminate to the general public the state of development of social robots. Therefore, the validation of the different stages of the development of our social robots will be performed in public performances. Taking into account the difficulties that robots have to adapt to different environmental conditions, this is not an easily attainable goal, but since social robots are intended to socially interact with humans, this seems the best way to evaluate the appropriateness of the work implemented.

## 1.2 Robotic platforms

The research activity described in this dissertation has been carried out mainly as an extension of the investigation conducted in the RSAIT<sup>4</sup> research group. RSAIT, founded around year 2000, is a small research group that develops its work at the Faculty of Informatics located in Donostia-San Sebastián, University of the Basque Country (UPV/EHU).

RSAIT owns a heterogeneous set of robots with which we can experiment and empirically evaluate the research done. Two different types of robots have been used in the present research work: several wheeled mobile platforms and two humanoid robots.

The mobile platforms were mainly designed to do research in indoor navigation techniques, but they were adapted with time to make them more suitable for interaction with humans.

*MariSorgin* is our heirloom robot, a synchro-drive robot that dates from 1996. It is a B21 model from Real World Interface provided with a ring of ultrasounds, infrared and tactile sensors for obstacle avoidance. In addition, a Hokuyo URG-30 laser, a Kinect sensor, a Heimann thermopile and a touch screen have been placed on top of the enclosure (see Figure 1.2(b)).

*Galtzagorri* and *Tartalo* (see Figures 1.2(d) and 1.2(c)) are two differential robots marketed by Omron Adept MobileRobots<sup>5</sup>. The former is a Pioneer-3DX model with a Leuze RS4 laser scanner, while the later is a PeopleBot platform with a Sick LMS200 laser sensor mounted on top of its base. For interaction purposes,

<sup>4</sup><http://www.sc.ehu.es/ccwrobot/seccion/home/lang/en>

<sup>5</sup><https://www.adept.com/home/?region=eu>

both are provided with a ring of ultrasound sensors, a Canon VCC5 ptz camera and a Gechic OnLap 13-inch touch screen added.

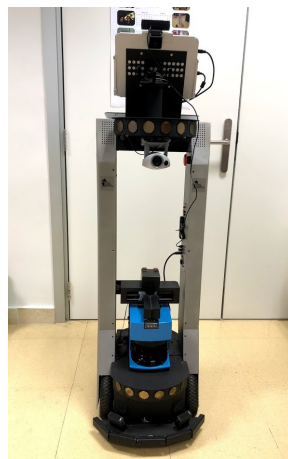
On the other hand, *Kbot* (see Figure 1.2(a)) was designed and developed by Neobotix<sup>6</sup> in 2004 to act as a tour guide at the Eureka Museum of Science in San Sebastian. It was put away in 2006 after breaking down, and was kept in a warehouse until RSAIT inherited it in 2015. The differential drive was repaired, and some components were removed giving it a more suitable morphology. The computer capability was substituted with a Zotac's Zbox mini PC and a Kinect sensor, and a smaller touch screen were mounted on it.



(a) Kbot



(b) MariSorgin



(c) Tartalo



(d) Galtxagorri

Figure 1.2: RSAIT's wheeled mobile robots

Regarding the humanoid platforms, *NAO* is a well-known humanoid biped robot developed by Softbank Robotics<sup>7</sup> designed to work on social capabilities. It has a height of 58 cm and has 25 DoF. *NAO* has 2 CMOS video cameras, full-colour RGB

<sup>6</sup><http://www.neobotix-robots.com/mobile-robots-overview.html>

<sup>7</sup><https://www.ald.softbankrobotics.com/en/robots/pepper>

LEDs placed in the forehead, eyes and ears, 4 omnidirectional microphones, and 2 loudspeakers in its head. It is also equipped with 2 sonars in its chest to detect obstacles, 8 FSR (Force Sensitive Sensors) in its feet, 2 bumpers at the front of each foot for collision detection and several tactile sensors located on its forehead to receive tactile input through touch (see Figure 1.3(a)).

Finally, *Pepper* is a humanoid robot, also developed by Softbank, with a human-like torso that is fitted onto a wheeled platform supplied with 3 omnidirectional wheels, with a height of 120 cm and which has 20 DoF. It is provided with full-colour RGB LEDs placed in forehead, eyes, ears and shoulders, 4 omnidirectional microphones, 2 loudspeakers, and 3 cameras (two RGB cameras and one 3D camera) placed in its head. It is also equipped with an inertial measurement unit, 2 ultrasound transmitters and receivers, 6 laser sensors and 3 bumpers placed in its base. Pepper also has tactile sensors in its hands and forehead (see Figure 1.3(b)).

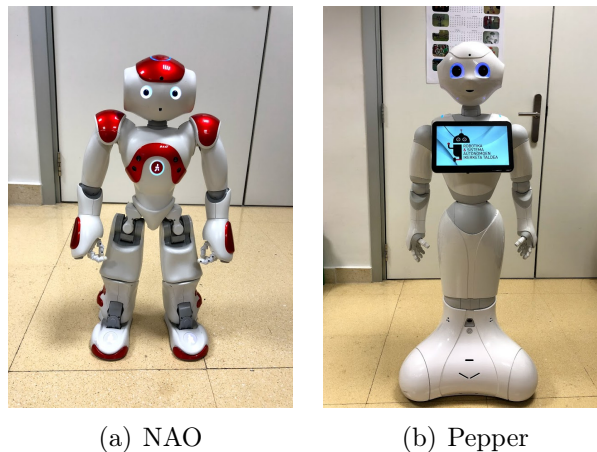


Figure 1.3: RSAIT's humanoid robots

Standardising the software for application developing enables the possibility of testing in different platforms the algorithms and behaviours being implemented with just small modifications, which in turn lightens the burden of rewriting code and adapting it. ROS<sup>8</sup> (Robot Operating System) is a well-known framework in the robotics community used for robot software development, which also provides operating system-like functionality on an heterogeneous computer cluster. ROS is a modular system that combines drivers and algorithms (such as navigation algorithms, control algorithms for robotic arms, etc.) to create robotic applications. Modules are named nodes in ROS and nodes communicate via topics or services following a publisher/subscriber protocol.

The first step before starting with this research work was to get all the robots running in ROS. Tartalo and Galtxagorri's basic drivers were already available thanks to the ROS community<sup>9</sup> whereas Kbot's and Marisorgin's drivers had to be implemented from scratch. Regarding the humanoid robots, both share a software

<sup>8</sup><http://www.ros.org/>

<sup>9</sup><http://wiki.ros.org>



named NAOqi<sup>10</sup> which was designed so that modules can be run independently across multiple machines and robots. Each module has an external API (i.e., functions and parameters) that other modules can call. The ROS community also offers a package named *naoqi\_driver* that wraps the needed parts of NAOqi library and make available the control of the robot in ROS.

After the setup step all RSAIT robots could be controlled using ROS, i.e we had a standard tool to uniformly use this society of heterogeneous robots.

### 1.3 Structure of the document

This document brings together the most relevant publications related to the research work carried out in this field. Those publications describe more in detail the work done and the results obtained during the experiments. They have been published in international journals and conference proceedings and therefore, they are available to the scientific community. We hope that the work carried out in the context of social robotics will be useful for future researchers.

This report has 5 parts which are divided into a total of 12 chapters. Chapter by chapter, the contents are organised as follows:

**Part I** is an introductory part that is composed by a single chapter.

**Chapter 1** includes an introduction to this research work, the motivation to start this project, the goals and the structure of the document.

**Part II** collects the research work related to the first steps made for social interaction with robots. Here, we start experimenting with different ways to communicate with robots and we propose two different applications: guide robots and a Natural User Interface (NUI) for entertainment.

**Chapter 2** provides an introduction to service robots and tour guide robots showing some social interaction capabilities (including interaction through GUIs among others), and it explains concisely the main concepts of robot navigation. At the end of the chapter, the GidaBot application, which allows our heterogeneous team of robots to act as tour guides in multi-floor buildings, is presented.

**Chapter 3** introduces robot teleoperation and its main applications, together with a review of the most relevant works related to robot operation using other NUIs (apart from GUIs). It also presents a body motion imitation interface that includes robot arm control and gesture-based robot locomotion commanding, and a speech-based commanding interface.

**Part III** brings together the work done in relation to the core of this dissertation, the BertsoBot system. It proposes a framework for socially interactive robots acting as verse-makers in the context of *bertsolaritza*.

---

<sup>10</sup><http://doc.aldebaran.com/2-5/naoqi/index.html>

**Chapter 4** gives an introduction in social robotics, including the most important features of social robots and some possible applications. It also proposes *bertsolaritza* as a showcase to disseminate the state of the art of social robots to the general public, and describes the ecological niche of the Bertsobot system.

**Chapter 5** defines the basic behaviours that constitute the Bertsobot's architecture. The developed modules that endow the robots with verbal communication, perception of the environment and non-verbal communication capabilities are also described in this chapter.

**Chapter 6** addresses the body posture recognition and action selection mechanisms developed to increase the autonomy of Bertsobot.

**Chapter 7** presents the two emotional behaviours developed for our social robots. The first one is related to the emotional response (according to the audience feedback) that the robot must show after singing a verse. The second one refers to the expression capability of the robot whilst talking.

**Chapter 8** presents the talking gestures generation system developed using Generative Adversarial Networks, and how it is coupled with emotional behaviour.

**Chapter 9** summarises the developed control framework and describes the evolution of the system through the different public performances carried out.

**Part IV** concludes this report.

**Chapter 10** gives the reader a general vision of the issues addressed by socially interactive robots. It summarises the key contributions and the lines of research that could be tackled in the future.

**Part V** collects all the publications related to the work performed in this research work.

**Chapter 11** groups the publications related to Part II

**Chapter 12** includes the publications related to Part III

## Part II

# First steps for social interaction with robots



# Chapter 2

## Interacting with tour guide robots

The evolution of robotics has always been linked to the human social needs. Robots have been used in factories since the 1960s helping to build products and relieving humans in performing dangerous or repetitive tasks. However, this situation is changing. The robot market, which has been increasing for decades in industrial applications, is now growing drastically in a wide range of service applications. These applications address the challenges of service robots capable of working in dynamic, uncertain, and uncontrolled environments alongside humans without being a hazard.

The coexistence between robots and humans is becoming a reality. Robots performing simple housework, transporting people, and even carrying out care tasks is no longer an unimaginable picture. But it is also a complicated scenario, with several problems to be considered. This scenario requires robots with certain survival capabilities [172]:

- Cognition: the robot's ability to perceive, understand, plan, and navigate in the real world.
- Manipulation: precise control and dexterity for manipulating objects in the environment.
- Interaction: the robot's ability to interact with humans, including support for verbal and non-verbal communications, observing and copying human behaviour, and learning from experiences.

This chapter presents the development of a system of heterogeneous robots collaborating as guides in multi-floor environments, which we have called GidaBot. Gidabot system includes some of the capabilities mentioned above; the system enables individual navigation and robot communication for cooperative guiding tasks in different floors, and incorporates a GUI that has been designed to facilitate the operation of the robot and improves the interaction with the user. The user can interact with the robot through the GUI we have developed, and the robot communicates with the user in return by using both the GUI and voice. Concerning the structure of this chapter, first, service robots and some possible applications are discussed. Next, we will introduce several tour guide robots and their interaction capabilities. After that, the concept of robot navigation and what a robot needs to

know in order to navigate in an unstructured environment will be described. Finally, the conclusions and our contributions in this field are presented.

## 2.1 Service robots

As mentioned before, robots are no longer limited to industry, they are progressively spreading to other domains (e.g. urban, social and assistive domains). Nowadays, research in robotics aims to develop socially interactive robots [55] that are going to live in our homes, workplaces, offices, etc. This type of robots pose two main challenges: on the one hand, they must be able to perform tasks in complex and unstructured environments, and on the other hand, they must naturally interact with humans.

Robots that aim to assist and/or perform useful tasks for humans are named service robots [81]. The idea of using robots in our daily lives is an inspiring research in the field of robotics. Service robots developed for different purposes can nowadays be found in some restaurants, offering a coffee to clients [124]; in hospitals, transporting critical patients to the surgery room [169]; in warehouses, moving material from location to location within distribution centres [173]; in retail stores, guiding the customers to the products of their choice in a shopping mall [70]; or even in museums, guiding visitors through the different rooms of the museum [149]. In an attempt to classify the different service robots that currently can be found in real environments, the International Federation of Robotics (IFR) proposes a list ordered by category and type of interaction [81]. On the one hand, there are those robots related to personal/domestic use, and on the other hand, those for professional use.

Despite the variety of already commercially available service robots, most of them have little presence in our lives yet. The most well-known exception is the iRobot's Roomba vacuum cleaner [61], which was designed to carry out the domestic task of cleaning the floor of our houses (see Figure 2.1(a)). Even though it is able to speak, its interaction capabilities are limited to basic interaction through the smartphone application or the robot's base buttons. Paro is another popular service robot, mostly used for therapeutic purposes with the elderly [112]. By interacting with people through its tactile and audition sensors, Paro responds as if it was alive, moving its head and legs, making sounds and imitating the voice of a real baby harp seal (see Figure 2.1(b)). Assistant-companion robots are another example of service robots with higher interaction capabilities than those kind of robots aforementioned, and that are currently gaining attention. Robots like Jibo [84] (see Figure 2.1(c)) or Buddy [24] are voice-assistants inside a robot body with high verbal-communication capabilities. They show some capabilities such as talking, recognising different users, answering some questions, etc., that make them more interactive. And although they are evolving little by little, at the present time they are just a seed of what humans expect of a friendly social companion robot.

Definitely, the level of interaction between robots and humans mostly depends on the task that the robot has to perform. A vacuum cleaner robot does not need the same interaction capabilities that assistant robots or guide robots do. Of course, not

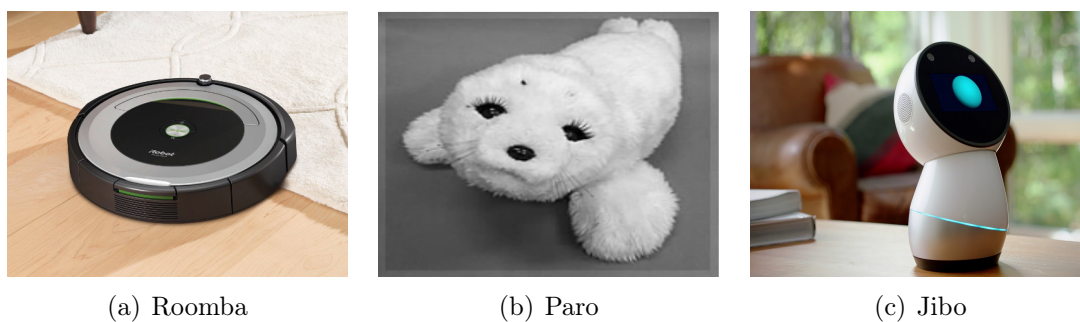


Figure 2.1: Several examples of service robots

all service robots have to be highly socially interactive, but in one way or another, they must be able to interact with humans.

## 2.2 Tour guide robots

Robots acting as tour guides is one possible application for service robots. Travelling and exploring new cities or small towns is one of the things that people like most. Museums, churches, and castles are places that tourists want to visit. However, exploring places with the help of a guide can be advantageous. Robots can help or replace human guides in these tasks, taking visitors from one place to another of the building, avoiding obstacles, and providing information about the place that the visitors have commanded.

The literature review reveals several instances of tour guide robots. Minerva [156] is very likely the first robot that acted as such in the Smithsonian's National Museum of American History in Washington, and by far the most cited one. Minerva interacts with people using a combination of its voice and facial expressions, and is commanded by the visitors through the touch-sensitive screen mounted on robots's back, with which they can select the tour they want to perform. In [134] the navigation capabilities of CoBot are evaluated while acting as a guide through a cooperation between the visitor and the robot, helping each other to fulfil the task. Its dialogue capabilities enable the robot to answer to task-related requests, and it can ask for help. The visitor also can interact with the robot through the GUI (displayed in the laptop screen) to help it localise in the map. More recently, kTBot, a robot that performs guided tours was designed, built and set up at the Eureka Science Museum of San Sebastian by Tekniker-IK4 [149]. The robot is able to interact with humans, interpret and understand their guiding requirements and plan the path to the destination. Visitors can choose between different points of interest to visit through the GUI showed in its touch screen located in its back. Some authors defend the need of humanoid platforms for social interaction with the visitors. Robovie [142] and Robotinho [54] are two humanoid robots with verbal communication and human-like body expression capabilities that act as guides at the Osaka Science Museum Exhibit and the Deustches Museum of Bonn respectively. Surprisingly, they are not provided with a touch screen to interact with visitors.

The research works aforementioned are limited to single robot navigation systems, and most of them perform the guiding tasks in a single floor of the building. Instead, references to multi-robot guiding systems are few. Trahanias et al. [159] present a different approach in which a group of robots can perform on-site and web-based tours. Its on-board interface allows the visitors to select tours and send the robots to different locations, while the web interface adds more characteristics, such as the option to teleoperate the robots from distance. In that vein, Hristoskova et al. [78] propose a distributed collaboration between two CoBots acting as guides. In addition to the characteristics and capabilities mentioned before, these robots share profiles and tour information with the aim of automatically exchanging the group members.

The problem of navigation becomes more complicated when the robot needs to travel to more than one floor. A possible solution to the multi-floor navigation problem is the use of a single robot that utilises lifts to navigate through different floors, as the robot Charlie [160] is able to do. A similar approach extended to multiple robots, also using elevators, is proposed in the GuideBot tour guide [104] and the BellBot hotel assistant [105] systems. In both systems, robots show some social capabilities, such as speaking and facial expression skills, and visitors and employers can interact with the assistant system through three types of GUIs: the on-board interface is used to interact with the robot and the overall system; the guest interface is located in the rooms allowing the user to request different services, such as guide the user to some place or bring her/him snacks, etc.; and the administrator interface that includes tools to see the location of all the robots in the building maps, monitor the state of the robots, list of tasks assigned to them and even monitor the on-board sensors. But entering lifts may be dangerous for robots; depending on the security measures, the size of the gap on the floor, the geometry of the robot, and especially the drive system, it may not be appropriate for the robot to use the lift. Moreover, robots that get into lifts are supposed to have the necessary abilities to interact with the lift interface, from inside and outside, to execute precise actions. The lack of proper actuators can be overcome by interacting with humans as CoBot does [134][165]. This symbiotic collaboration approach has been further expanded to a homogeneous team of up to 4 robots that are also able to perform delivery tasks [164].

An alternate solution would be the use of Internet of Things (IoT), a currently fashionable topic that aims to connect different devices via network enabling them to exchange data. In this way, robots would be connected to the network, being part of smart buildings, and they would remotely control a lift in order to move from one floor of the building to another. However, IoT technology is still not mature enough, and for now few buildings are equipped with the required measures.

## 2.3 Robot navigation

A robust guide system mainly relies on robust navigation capabilities. For any service robot it is fundamental to be able to safely and accurately navigate in its



environment, and so it is for guide robots. Robot navigation implies that the robot is able to determine its own position in the environment and move towards the goal location. The most well-known classical solution to the navigation task is the probabilistic approach. The key idea in probabilistic robotics is to explicitly represent the uncertainty in robot perception and action using probability theory [157]. Within the probabilistic robotics field, navigation consists of answering three questions:

1. Where am I?
2. Where are the other places with regard to my position?
3. How do I get there?

The answer to these three questions is performed through three fundamental processes: the construction and interpretation of environmental representation (maps), self-localisation, and trajectory planning.

The construction of the map is usually made during a learning phase of the navigation, called mapping. In this step, a model of the environment is acquired by the robot using its sensors. Solution algorithms to the problem of SLAM (Simultaneous Localisation and Mapping) take into account the uncertainty in the robot localisation while building the map. SLAM has been and it still is one of the most successful research area in the field of robot navigation and more specifically, in probabilistic robotics. It offers different approaches for building different map representations and Occupancy Grid Mapping is commonly used nowadays for indoor robot navigation.

During the localisation step, the robot establishes its own position and orientation within the map. Again, there are several probabilistic approaches to robot localisation that cope with the uncertainty associated to odometry. One of the most well-known localisation algorithm is the Adaptive Monte Carlo Localisation (AMCL) method. It uses a particle filter to track the pose of a robot against a known map. Other common approaches are the Markov Localisation algorithm, that is the straightforward application of the Bayes filter to the localisation problem, and the Kalman filter algorithm, a technique for filtering and predicting linear Gaussian systems that represents the belief through a multivariable Gaussian function.

Finally, in the last step, the planning of the trajectory to go from the current robot position to the goal location is calculated and executed. The trajectory planning problem can be divided into two sub-problems: global planning (path planning) and local planning (obstacle avoidance). Most common approaches for trajectory planning are probabilistic roadmap methods, grid based algorithms (e.g. A\* greedy algorithm) and reward-based algorithms (see [145] for a more detailed review).

ROS provides a navigation stack initially developed for the PR2 robot by Willow Garage [109] that has been adapted for many robots<sup>1</sup>. This navigation stack

---

<sup>1</sup><http://wiki.ros.org/navigation/RobotsUsingNavStack>

offers tools for constructing a global map of the robot's environment by means of a SLAM technique that uses a RAO-Blackwellized particle filter in which each particle represents an individual map of the environment. It also uses an adaptive technique to redistribute particles across high probability regions of the probability density function and avoid the problem of particle depletion<sup>2</sup> techniques. Besides, robot localisation during navigation is maintained using the AMCL algorithm. Together with the map and robot localisation mechanisms, the navigation stack needs a planner to find and select the path to be followed by the robot. ROS allows the user to configure the stack by choosing, among several planners, the one that better fits to the robot/environment system. The default navigation function makes use of Dijkstra's algorithm for planning purposes.

## 2.4 Contributions and conclusions

Our contribution related to the topic of this chapter is a system of heterogeneous robots collaborating as guides in multi-floor environments. We have developed a system, which we have called GidaBot, that enables robot communication for cooperative guiding tasks in different floors, and allows individual navigation in each floor at the same time. The system is designed to operate in buildings where robots can not move from one floor to another. Its robustness has been tested using four real robots (Tartalo, Kbot, Galtxagorri and Marisogin) at the Faculty of Informatics in Donostia-San Sebastián, one on each floor of the building.

The first step of the process to reach the actual state of the GidaBot system was to setup all the robots, to develop the missing drivers and to establish a uniform configuration for all of them. ROS has provided us the opportunity to set the same programming and control environment to standardise our society of robots. Having all robots "standardised" with ROS next step was to endow them with navigation capabilities. The multi-floor guide system developed in this work makes use of the ROS navigation stack, adapted to each of the platforms involved in the system.

Kbot was the first of our robots to become a robot guide. The tour-guide system implemented in Kbot allows the robot to perform guiding tasks in the first floor of our Faculty. The system relies on the ROS navigation stack and a Qt<sup>3</sup> based GUI; ROS navigation tools have been used to build the map, localise the robot within the map during navigation, and to plan the trajectory to move towards the goal location, while the Qt-GUI permits the user to interact with the system by choosing the desired goal in the map displayed in the robot's touch screen.

The research work related to the single robot navigation system aforementioned is collected in the following publication:

- **Standardization of a Heterogeneous Robots Society Based on ROS.** I. Rodriguez, E. Jauregi, A. Astigarraga, T. Ruiz, E. Lazkano. In A. Koubaa (Ed.) Robot Operating System (ROS), Volume 1, 2016, pp. 289-313. Springer.

---

<sup>2</sup><https://openslam-org.github.io>

<sup>3</sup><http://wiki.ros.org/IDEs#QtCreator>

The single robot navigation setup carried out in Kbot has been used as a base for developing the interfloor multi-robot guide system. The previous system limited the operation of the robot to a single floor, thus the next goal was to extend it to be able to allow tours all along the different floors of the building. Four robots have been used to cover the four floors of our Faculty, which are communicated to perform cooperative guiding tasks. The interfloor multi-robot guide system provides two different operation modes: single target mode and tour mode.

In single target mode, the robots must cope with different situations: the simplest case is when the user and the goal are in the same floor; only one robot will guide the user from the beginning to the end of the navigation. However, when the user and the desired goal are in different floors, two robots are involved in the guiding task; the first will drive the user to its floor meeting point (the lift or staircase), while the second one will be waiting for the user in the goal floor's meeting point to solve the remaining path to the goal location. Each robot has its request queue, and pending goals are processed in the same order they are requested.

Often, when welcoming visitors it is interesting to follow a predefined sequence of locations, i.e. to offer guided tours. Those tours should be adapted to visitors' profiles, giving priority to some locations and focusing on the most interesting location for them. For this purpose, the system allows to create tours as a collection of location goals in the desired sequence. Tours are saved in a local directory and can be edited. New and edited tours are automatically shared among all the available robots involved in the tour.

Extending the system to allow guided tours among different floors also requires improvements in usability and intuitiveness of the GUI to satisfy the users' needs. The previously developed GUI was renewed, new options related to the operation modes (single target and tour modes) were added, and actually information about all the robots that are part of the system is displayed. It also shows informative messages to the user that are supported by the speech. The speech system employed in GidaBot is the same developed for the NAO robot guidance system described in the next chapter (see Chapter 3.3). Figure 2.2 shows the renewed aspect of the interface.

In order to make the system work as desired, each robot has to inform the others, on the one hand, about user's requests and, on the other hand, about its state (current location and navigation state). They exchange different types of messages concerning goal descriptions, tour description, robots position, and pending requests.

Different experiments have been carried out to evaluate the robustness of the Gidabot system. Experiments have been performed in both simulated and real world environments (see videos <sup>4,5</sup>). The application has been used for several times since 2015 in the open door event held at our faculty every year. About 100 candidate students come every year to visit the facilities, and divided into groups they follow a tour in which our robots show them the most interesting rooms and places of the faculty. Figure 2.3 shows Kbot and Marisorgin making a guided tour with bachelor

---

<sup>4</sup><https://www.youtube.com/watch?v=fER54Me-qcU>

<sup>5</sup><https://www.youtube.com/watch?v=i1UtxrGieks>

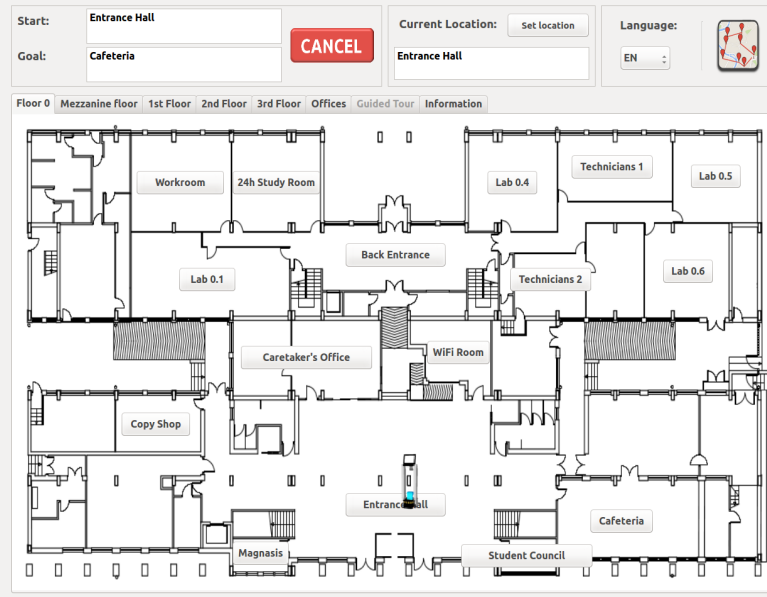


Figure 2.2: Main view of the renewed GUI

students in the first and fourth floors of the faculty, respectively.

The research work related to the interfloor multi-robot guide system described before is collected in the following publication:

- **GidaBot: a system of heterogeneous robots collaborating as guides in multi-floor environments.** O. Parra, I. Rodriguez, E. Lazkano, T. Ruiz. International Journal of Advanced Robotic Systems (IJARS), 2017 (Submitted).

To conclude, we must say that robots performing tasks in human settings must be able to adapt to the environment. Navigation is a basic ability that all mobile robots must be endowed with. They should be able to autonomously explore the environment in order to obtain the information necessary to create a map of the surroundings, but also able to localise themselves in it.

Concerning service robots, there are many applications in which they can be useful to assist or help humans performing different tasks. This type of robots are very useful for guiding purposes. Endowed with the appropriate capabilities, such as verbal-communication and facial expression skills, they can take visitors from one place to another while offering information (through speech or touch screens) about the location that the visitors want to know.

Albeit the GidaBot system was set up to solve the multi-floor navigation problem of a specific environment, our Faculty, the system can be adapted to a different building and robot configuration. It would be necessary to map the new environment and to setup the GUI with the interesting places. Besides, GidaBot's system robustness depends on the performance of two building blocks: navigation behaviour and wireless communication. The ROS navigation stack offered us the basic tools to setup the navigation capabilities of our robots. It showed to work well in spite of



(a) Kbot showing the students the first floor of the faculty



(b) Marisorgin showing the students the fourth floor of the faculty

Figure 2.3: Kbot and Marisorgin making guided tours in the faculty

some bizarre “turning on the spot” tendency when trying to relocalise after getting lost. On the contrary, setting up robust wireless communication in a public building has shown to be a big burden and the obtained behaviour still needs refinement to eliminate break downs during a guiding task.

Regarding the interacting methods, it is undeniable that verbal communication is more natural and intuitive. But GUIs are becoming more and more present in our dairy lives thanks to mobile devices and they can display a window of the options the system offers in a single view.

## Chapter 3

# Interacting with a humanoid robot through Natural User Interfaces

Since social robots directly interact with people, finding “natural” and easy to use user interfaces is of fundamental importance. In the previous chapter we analysed some of the social capabilities that service robots show, such as talking, speech recognition, etc. We have also seen that in addition to those skills, in the context of tour guide robots, the most usual way for the users to command robots is through a GUI displayed in robots’ touch screens.

Natural user interfaces are an emerging technology that enables users to interact with robots through natural means, such as body and voice, eliminating the traditional interfaces like the keyboard to command the robot. This chapter presents a remote control (teleoperation) system that aims to humanise the way to operate a humanoid robot. It is composed by two natural interfaces: a body motion imitation interface with which the user can control the robot arms and command the robot through body gestures recognised using the Kinect sensor; and a speech-based commanding interface. Regarding the structure of this chapter, first remotely controlled robots and their main applications are introduced. Next, a review of the principal works related to robot teleoperation using motion capture devices is presented together with a review of robot control through body imitation. Finally, the conclusions and our contributions in this field are presented.

### 3.1 Remotely controlled robots

Teleoperation is the term used in research and technical communities to allude to operating a system remotely [60]. In other words, it refers to the action in which a slave manipulator reproduces faithfully the movements of a master manipulator, controlled manually by a human operator. Thus, a teleoperated system is a system based on a master-slave communication model that in addition to extend the human capability of manipulating an object at distance, it provides the operator with the feedback of the action performed in the remote location.

The concept of teleoperation is commonly associated with robotics and mobile

robots, although it can be also applied to a wide range of areas, such as entertainment systems, industrial machinery, etc. From now on, when mentioning teleoperated systems we will be referring to the teleoperation of robots.

In a typical scenario, a remote controlled robot is operated by a user that sends commands to the robot from a remote centre, and supervises the performed motion by receiving feedback from its sensors. This type of systems are conceptually divided into two parts: local area and remote area. The local area represents the place where the human operator (master) and all the necessary elements to communicate with the remote site and receive the feedback from it are located. On the other side, the remote area comprises the environment to be manipulated and the robot (slave), equipped with the required sensors to perceive the environment. Figure 3.1 shows a general overview of a robot teleoperation system.

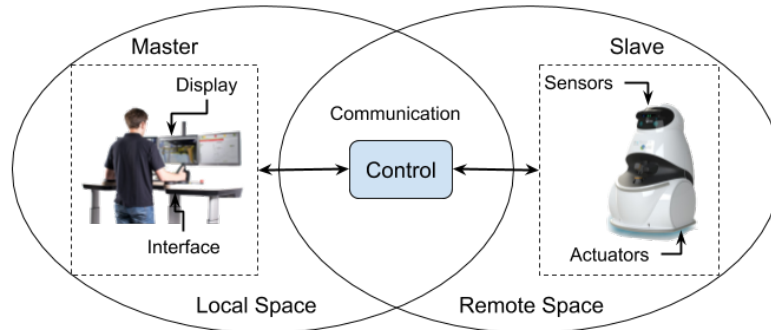


Figure 3.1: General overview of a robot teleoperation system

We have been using different tools to increase our manipulation abilities since ancient times. Tools like simple sticks have allowed us to reach objects that were at a distance, or others like clamps have helped us to manipulate dangerous pieces. Looking back in history, we find that research activities about remote manipulation of objects were born in laboratories of nuclear industry, however, over the years its applicability has been extended to other sectors of society. Some of the most significant fields of application of teleoperation are listed below:

- Space exploration: the difficulties of sending humans safely to remote and dangerous environments make teleoperated vehicles the best resource to explore and collect information in the space. The most known example of exploration rovers are NASA's *Spirit* and *Opportunity*, which were developed for searching evidences of past water activity in Mars [148][23]. These rovers were not directly teleoperated. Instead, they received a sequence of commands every Martian morning. Besides, humanoid robots have also been sent into space; Robonaut is currently operated inside the International Space Station (ISS) and can be teleoperated to mimic the motions of a crew member wearing a 3D-view device, a vest, and specialised gloves [45].
- Rescue: Search and Rescue Robots (SAR) are a tool aimed to help people. The overall goal of teleoperated rescue robots is to explore unknown disaster scenes



while searching for victims in situations, like earthquakes, urban disasters, or explosions, where humans can not fit or operate. They have been used in the rescue and recovery operations of many past devastation, such as the World Trade Center (WTC) catastrophic damage [116] and Tohoku earthquake in Japan [71].

- **Surgery:** remote surgery allows surgeons to perform precise operations from remote locations. A robotic surgical system requires advanced communication technologies that allow the surgeon to operate the patient through the control of a robotic arm and receive feedback from a sensory system in real time. Da Vinci is one of the most well-known commercial surgical system, designed to facilitate complex surgery using a minimally invasive approach [14]. It is commonly used for prostatectomies, and increasingly for cardiac valve repair and gynaecology surgical procedures. A newer remote surgery robot generation is the robot Versius, which uses five self-contained robotic surgical arms that are part of a modular system adaptable to the surgery [31].
- **Telepresence:** the aim of a telepresence system is to replace a human presence with a robot at a remote location, and in turn, to make the operator feel that she/he is present at the remote site. A telepresence robot is typically composed by a touch-screen and a wheeled mobile base equipped with vision and sound technologies. These sort of robots are used for different purposes, such as for elderly healthcare [92], education [152], or home-rehabilitation [25]. More sophisticated and with a very realistic human appearance are Ishiguro's geminoids, which are used by its creator to replace him in meetings and conferences [136].
- **Entertainment:** entertainment robots are designed for recreational purpose and are increasingly common to find in our daily life. There are many types of entertainment robots which have been developed for different purposes. A prominent example are drones, that are used for different applications, such as remotely controlled mission games, to compose music or in live performances [89]. Social robots like Cozmo<sup>1</sup> are also designed to entertain the user. Cozmo plays with its cubes challenging the user with its favourite games of speed and skill, but it can also be used in explorer mode (remotely controlled) to guide it through its environment.

## 3.2 Natural user interfaces for remote controlled robots

Despite the fact that touch screen-based GUIs improve the sense of control of the robot over those GUIs that need a mouse or keyboard to interact with the system, there is still a barrier in the communication between humans and robots. Together

---

<sup>1</sup><https://www.anki.com/en-us/cozmo>

with advances in technology there has also been progress in human-machine interfaces looking for a greater sense of control of the machine. Consequently, there is a growing demand to create more immerse user interfaces that take full advantage of modern technologies allowing users to act with the machine in a more natural way. That is precisely the aim of natural user interfaces: to allow users to engage with machines in a similar way they would interact with the real world through using body movements, hands or even voice [126].

Each environment and each task to be performed require a different type of teleoperation interface. And naturally, the interface varies depending on the type of robot; the teleoperation of a wheeled robot differs from that of a humanoid robot, which has more degrees of freedom to be controlled. According to Fong et al. [60] there are four categories for teleoperation interfaces: direct interfaces enable the operator to control the robot via hand-controllers (e.g. joystick, keyboard, or touch screen) obtaining the feedback from the robot's cameras; multimodal/multisensor interfaces allow the operator more complex control modes (individual actuator, coordinated motion, etc.) and/or integrate information from various sources in one view (text, visual, sound, etc.); in supervisory control interfaces the user is able to send high-level commands, monitor the remote scene and get a diagnosis; and novel interfaces have special input methods (e.g. gestures) or input-output devices (haptics), or used new display systems (virtual reality).

Early approaches for humanoid robot teleoperation were mostly based on a control by direct and multimodal/multisensor interfaces, which captured user intentions through the use of joysticks, buttons and keyboards [144], or by the use of Graphical User Interfaces (GUIs) [151]. Nowadays, research is focused on developing more advanced interfaces that permit a more natural interaction between the operator and robot. Motion capturing devices are very suitable for that purpose, particularly for commanding humanoid robots. They enrich the teleoperation allowing the user to operate robots by means of gestures or by imitating whole body motion. In addition, the advances in software development, such as modern tools for voice recognition, make possible a more instinctive control of the robot. In this way, the operator can exchange information with the robot by verbal communication, giving it orders while interacting in a natural way.

### 3.2.1 Teleoperation by motion imitation

In recent years, many new motion capture products have come onto the market, ranging from depth cameras to full-body motion capture suits. This type of devices have opened new research paths in robotics that led researchers to develop systems which help to improve humanoid locomotion [125] or allow robots learning from human demonstration [5] among other things.

Motion capture devices are also very suitable for robot teleoperation, specially for humanoid robots that require advanced control due to its amount of degrees of freedom. Gesture-based teleoperation is increasing adepts specially due to availability of cheap depth cameras like Microsoft's Kinect sensor. Such devices allow the operator to drive the robot in a more natural way, by means of simple communi-

cation channels, like gestures. Several works can be found related to gesture-based teleoperation systems using depth camera images: Tara et al. [154] proposed a sign language recognition system for robot teleoperation in which they acquire hand gesture information from depth images. Du et al. [46] also used the Kinect sensor in order to track human hand motions aiming to control a robot manipulator to perform pick and place tasks.

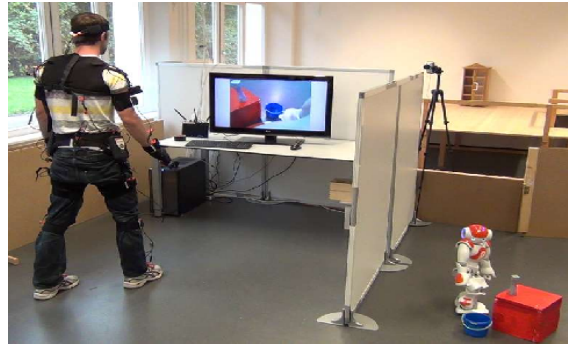
Real-time teleoperation of humanoid robots by imitating body motion (full-body or some parts) is another active research area. Related work on this topic is abundant. For instance, Setapen et al. [140] used an active-marker motion capture system to teleoperate a NAO humanoid robot, with the aim of teaching the robot new motions. They applied inverse kinematic calculations for finding the mapping between motion capture data and robot actuator commands. Matsui et al. [111] applied a capture system based on body markers to measure the motion of both, a humanoid robot and a human, and then adjust the robot motion to minimise the differences, with the aim of creating more naturalistic movements on the robot. Song et al. [146] utilised a custom-built wearable motion capture system, consisting of flex sensors and photo detectors, to convert motion capture data to joint angles. Koenemann and Bennewitz [93] presented a system that enables a humanoid robot to imitate complex whole-body motions of humans in real time, ensuring static stability when the motions are executed and capturing the human data with an Xsens MVN motion capture system consisting of inertial sensors attached to the body (see Figure 3.2(a)).

The above mentioned methods are limited in the sense that the human needs to wear different types of sensors in order to interact with the robot. Moreover, the cost of the equipment is quite high in comparison with depth-sensing cameras. That is why researchers have become more interested in using depth cameras for humanoid teleoperation. Song et al. [147] proposed a teleoperation system using a Kinect sensor to capture human motion and control the actions of the Robonova robot. A more advanced system to teleoperate a robot with higher DoFs was proposed by Almetwally and Mallen [3]. Their technique also used the Kinect and allowed the user to move not only the arms but also to drive the robot. A similar approach to imitate the arms and head movement of the user was presented by Li et al. [102]. Finally, Ou et al. [119] propose a human imitation system for NAO using the Kinect sensor. The robot mimics the whole body motions made by the user in real time while maintaining the balance (see Figure 3.2(b)).

Any humanoid robot teleoperation system based on body motion or gestures imitation requires an interface composed at least by two main elements:

- A system that captures the motion or gesture made by the user.
- A control system that translates the captured motion to the robot space.

The main problem of humanoid robot teleoperation, in particular that based on motion imitation, lies on converting a movement from the capturing system space into the robot space. This conversion is known as inverse kinematics and usually has no solution. Thus, the system must provide an approximated solution using



(a) Koenmann et al.'s teleoperation [93]



(b) Ou et al.'s teleoperation [119]

Figure 3.2: Two examples of NAO teleoperation through motion capturing interfaces

kinematics equations, and in turn it must also guarantee the robot stability while imitating the human movements.

### 3.2.2 Teleoperation by verbal commands

Verbal communication should be a natural way of human-robot interaction. Humans feel more comfortable when the interaction with machines is through voice. Interaction by verbal communication requires that both, the interlocutor and the receiver use the same communication channel. In this way, the operator can give verbal commands to the robot, which the latter must understand, in order to perform the task associated with the order.

To serve people, it is necessary to develop an active auditory perception system for the robot that can execute various tasks in everyday environments obeying spoken orders given by a human and answering accordingly. Several systems have been developed that permit natural-language human-robot interaction. Foster et al. [62] proposed a human-robot dialogue system for the robot JAST, where the user and the robot work together to assemble wooden construction toys on a common workspace, coordinating their actions through speech, gestures, and facial displays. Wang et al. [168] introduced a human-robot speech system for teleoperating a humanoid mobile robot that can move around in environments, and perform physical tasks, such as searching objects. The speech control is event-based in order to avoid

communication delays. Gallardo and Poncela [66] presented a human-robot interface for command-based voice teleoperation of a Pioneer P2AT robotic platform. They also developed a user and command dependent acoustic model in Spanish for voice recognition commands.

Any speech based teleoperation interface should provide the user the possibility to control the robot giving voice orders. The interface receives the user's input in the form of voice command (captured from a microphone), and in response the interface interacts with the robot sending the corresponding commands. Two main elements are identified in an architecture for speech-based teleoperation:

- The automatic speech recognition system (ASR)
- The robot control system

The system also should feedback the operator when an instruction is not understood, and this feedback should also be verbal. For that purpose another element is required, a text-to-speech (TTS) system that converts the text to be said by the robot into speech.

Most speech-based teleoperation systems, including those mentioned before, make use of external tools in order to capture and recognise the speech, and also for converting the text into speech. It is worth mentioning some of the most known ASR systems, such as Microsoft's Speech Recognizer<sup>2</sup>, Sphinx<sup>3</sup>, Julius<sup>4</sup>, or Google's speech-to-text engine<sup>5</sup>.

On the other hand, with respect to TTS systems, there is much more variety, and some of them are even multi-language, such as Acapela<sup>6</sup>, Nuance<sup>7</sup>, or Google's TTS engine<sup>8</sup>.

### 3.3 Contributions and conclusions

In this chapter, we propose a real-time humanoid robot commanding system using two different natural user interfaces to enrich the interaction with NAO: a body motion imitation interface that includes the option of arm motion imitation and gesture-based robot locomotion commanding; and a speech-based commanding interface. Here we will describe each one separately, as if it were two different robot teleoperation systems, but both can be combined in a single one to control simultaneously the robot through speech and body motion.

The body motion imitation interface proposed in this research work, which is based on motion capturing using the Kinect sensor, allows the user to control not

---

<sup>2</sup>[https://msdn.microsoft.com/en-us/library/hh378380\(v=office.14\).aspx](https://msdn.microsoft.com/en-us/library/hh378380(v=office.14).aspx)

<sup>3</sup><http://www.sphinx-doc.org/en/master/>

<sup>4</sup>[http://julius.osdn.jp/en\\_index.php](http://julius.osdn.jp/en_index.php)

<sup>5</sup><https://cloud.google.com/speech-to-text/>

<sup>6</sup><http://www.acapela-group.com/>

<sup>7</sup><https://www.nuance.com/dragon.html>

<sup>8</sup><https://cloud.google.com/text-to-speech/>

only the arms of the NAO robot but also to navigate the robot. Regarding the control of the arms, our approach is based on imitation of movements, i.e. NAO replicates the operator's arm movements. For the purpose of imitation, first human skeleton is tracked, then data are suitably transformed to NAO's coordinate space, and finally, the gesture is executed by NAO. To transform the Cartesian coordinates obtained from the Kinect to NAO's coordinate space, a joint control approach is employed; human joint angles at shoulders and elbows are calculated and translated to NAO's coordinate space. On the other hand, for robot locomotion we propose to use a predefined set of gestures (based on body positions). The operator's body position defines the action she/he wants the robot to perform, for instance, to make the robot move forwards the user has to take a step forward, to make the robot turn right she/he has to lean the shoulder to the right, etc. Moreover, a GUI that shows the visual information obtained from both NAO's camera and the Kinect's camera has been developed, that facilitates to the user the control of the robot from distance (see Figure 3.3).

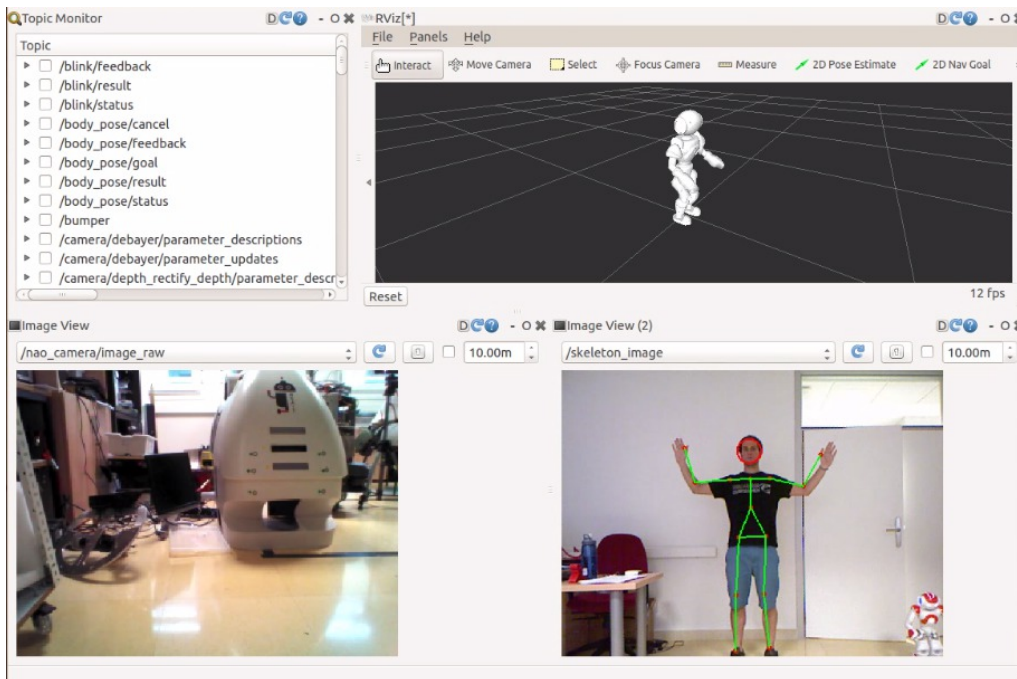


Figure 3.3: Teleoperation graphical user interface

The speech-based commanding interface presented in this research work to command NAO follows the three elements architecture aforementioned; it is composed by ASR, robot control and TTS systems. The order given by the operator is captured by the ASR system that uses two external tools to translate it into text; SOX<sup>9</sup> is employed to capture the audio and Google speech-to-text engine to convert it into text. Then, the action to be executed by the robot is determined according to the text obtained in the previous step; keywords that matches the list of commands

<sup>9</sup><http://sox.sourceforge.net/>

defined for the control of the robot are looked up in the sentence. If a match is found, the robot performs the movement corresponding to the received command, such as “Stand up”, “Move forward”, “Turn left”, etc. Otherwise, the robot says that it could not understand the order and asks the user to try again. AhoTTS [76] has been used to provide the robot with speaking capability. Its main task is to convert the text to speech. AhoTTS was mainly developed for Basque language, but it is also available for Spanish and English.

The research work related to the natural user interfaces for humanising the way to interact with NAO is collected in the following publication:

- **Humanizing NAO robot teleoperation using ROS.** I. Rodriguez, E. Jaurregi, A. Astigarraga, T. Ruiz, E. Lazkano. 2014 IEEE-RAS International Conference on Humanoid Robots (Humanoids), Madrid, Spain, November 2014, pp. 179-186.

Experiments in the laboratory have been conducted to evaluate the usefulness of the natural user interfaces for robot commanding proposed in this chapter. We have recorded two videos during the experiments carried out in the laboratory:

- The first video<sup>10</sup> shows how the operator commands the robot to transport an object from one place to another. The aim of this experiment was to test the arm and walking movements of the robot while performing a simple task in cooperation with humans.
- The second video<sup>11</sup> shows the guidance of the robot using speech commands. The system also gives a feedback to the operator when an instruction is not understood, and this feedback is also verbal.

The experiments carried out revealed three aspects that might be improved:

1. The lack of side view makes more difficult the guidance of the robot: the head control was afterwards added into the system in order to alleviate this problem. Currently the user can control NAO’s head movements by moving her/his own head. The robot’s head joint pitch and yaw angles are calculated just like in the joint control approach employed for the arm motion imitation.

Another experiment was carried out in order to test the full system, this time including the arm, head and walking control options (see video<sup>12</sup>). The following publication includes the full system:

- **Standardization of a Heterogeneous Robots Society Based on ROS.** I. Rodriguez, E. Jauregi, A. Astigarraga, T. Ruiz, E. Lazkano. In A. Koubaa (Ed.) Robot Operating System (ROS), Volume 1, 2016, pp. 289-313. Springer.

<sup>10</sup><https://www.youtube.com/watch?v=Toacwmm9OkU>

<sup>11</sup><https://www.youtube.com/watch?v=ynfNMgZjiVA>

<sup>12</sup><https://www.youtube.com/watch?v=EaFrgyZorFA>

Furthermore, the two natural user interfaces proposed in this chapter were used and tested in a public performance held in December 2014. NAO was invited to the “ScienceClub 2014”, an event that aim to disclose science and technologies to the society. In such event, NAO showed its imitation and verbal communication capabilities, mimicking the body motion the user performed and maintaining a simple dialogue with the presenter (see video<sup>13</sup>).

To end the chapter, some conclusions are extracted. Although teleoperation can be considered an ancient concept in robotics, there are still many situations in which robots require to be supervised and teleoperated by an operator. For instance, it provides the possibility of performing complex tasks in environments otherwise inaccessible or dangerous. Additionally, teleoperation can be used for robots to learn imitating human motion. Imitation is also a way of social interaction. A social robot must have the capability to imitate the agents around it. In a human society, people generally teach new skills to other people by demonstration; we do not learn to dance by programming, instead we see other dancers and try to imitate them. Hence, our artificial partners should be able to learn from us by watching what we do. In that vein, natural user interfaces make easier the way to interact with robots, allowing users to command robots through natural means like body motion and gestures, and even through voice.

The developed application was designed for entertainment purposes and it has been used several times as a such. But it must be taken into account that the tool itself can have many applications. We made one step forward and proposed the system as a rehabilitation tool. In the following publication a system that aims to monitor home rehabilitation exercises is presented. This system uses Machine Learning paradigms to classify human poses; a NAO robot is used to motivate the users to perform the rehabilitation exercises at home, and a Kinect sensor to measure the quality of the movements. NAO reproduces the exercises that have been taught through the body motion imitation interface presented before.

- **NAO Robot as Rehabilitation Assistant in a Kinect Controlled System.** I. Rodriguez, A. Aguado, O. Parra, E. Lazkano and B. Sierra. International Conference on Neurorehabilitation, Segovia, Spain, October 2016, pp. 419-423.

---

<sup>13</sup><https://www.youtube.com/watch?v=HKxe40-Qi6w>



## Part III

# A framework for socially interactive robots



# Chapter 4

## Socially interactive robots

Human behaviour has been studied from many perspectives over years. Psychology, neuroscience, or other disciplines like sociology, aim to study and understand the different aspects of human behaviour. Robotics also offers a complementary perspective on the study of human social behaviour. Understanding how we perceive and interact with others is currently a core challenge in robotics research.

Social robotics aims to provide robots with artificial social intelligence to improve human-machine interaction and to introduce them in complex human contexts [27]. The demand for sophisticated robot behaviours requires to model and implement human-like capabilities to sense, process, conduct high-level dialogue, learn, and act/interact naturally by taking into account emotions, intentions, motivations, and other related cognitive functions. And, of course, the ability to communicate through natural language and non-verbal signs is in the front line of research.

Verbal and non-verbal communication are therefore essential skills that any Socially Interactive Robot (SIR) [55] [59] must show. Naturally, speech plays a relevant role to convey messages or emotions, however facial expressions, voice intonations, or body expressions can disclose as much information as words. The design of a social platform has an important function when a natural HRI is intended. Physiological and biological studies conducted over the years describe how the design of robots affects the interaction between humans and robots [75], and the importance of developing robots with anthropomorphic (human-like appearance) features [47] for natural interaction. Both, anthropomorphic design and social skills with which robots are programmed, will help to increase the empathy and the acceptance level of robots. On the contrary, robots with extreme likeness to humans can elicit uncanny feelings of rejection and revulsion in observers [107].

In the last years, research in the field of social robotics has considerably grown up, and several robots showing verbal and non-verbal communication capabilities have been developed in this area. Robots endowed with social abilities have been used for different applications, for instance, as interactive teachers [65] and educational assistants [87], to support diabetes management in children [32], to assist elderly with special needs [21], or even as flyers delivery in shopping malls [141].

In addition to the applications mentioned before, social interactive robots can also be used for entertainment purposes. In the following section we will explain

and describe several examples of robots acting on stages.

## 4.1 Robots on stage

Entertainment is an area in which social robots can have high impact. Social robots can be used for different activities associated with amusement and which aim to hold the interest of an audience. Moreover, cultural Robotics, defined as the study of robots that participate in or create culture, is an emerging field that contributes to the advancement of social robotics [48], and several works make reference to robots participating in cultural activities. Fridin [64] proposes a robot that tells stories to kindergarten children. The humanoid robot employed in her work is able to show appropriate emotions through body expressions and voice intonations according to the story being told. Related to the art of music, Taheri et al. [150] present a robot with verbal capabilities for teaching music to autistic children, which also plays the xylophone and the drum. But robots can also dance. Augello et al. [12] describe a cognitive architecture for a humanoid robot that makes it able to create and perform dances driven by the perception of music. The humanoid robot also reacts to human mate dancers, tracking their face and listening the knocking sound they made with hands.

Theatre performances using robots show to be an appropriate showcase for disclosing the state of the art of social robots to the general public, and thus, to measure social acceptance of robots. Although everything is rehearsed beforehand, theatre offers an invaluable sphere to research and develop social behaviours in robots, to work and extend the expression of emotions and the natural communication among humans and robots [103][57]. Theatrical performances are also being used to evaluate HRI features [34] and to develop plausible scenarios for socially assistive robots [85]. The experiment made by Ogawa et al. [118] to measure the advantages androids might have as poetry-reciting agents is remarkable. A review of robot performances can be found in [117][106]. Little by little, robots are appearing in theatres motivated by researchers as a means, but also by artists [101].

## 4.2 Bertsolaritza

*Bertsolaritza*, the art of creating extemporary verses in *Euskara* (the language of the inhabitants of the Basque Country), is one of the manifestations of traditional Basque culture that is still very much alive. The Basque troubadours, named *bertsolari*-s, use this improvisational poetry context not merely to entertain but to discuss contemporary social, cultural, sexual, and political problems.

Events and competitions (see Figure 4.1(a)) in which the verse-makers have to produce impromptu compositions, named *bertso*-s, about topics or prompts are very common. A typical scenario involves an emcee (or presenter) suggesting a topic to the *bertsolari*, who must then, within the space of less than a minute, compose and sing a poem along the pattern of a prescribed verse-form that also involves a rhyme

scheme and a melody. And of course, the *bertsolari* must perform that verse, in front of an audience and without any musical accompaniment (see Figure 4.1(b)).

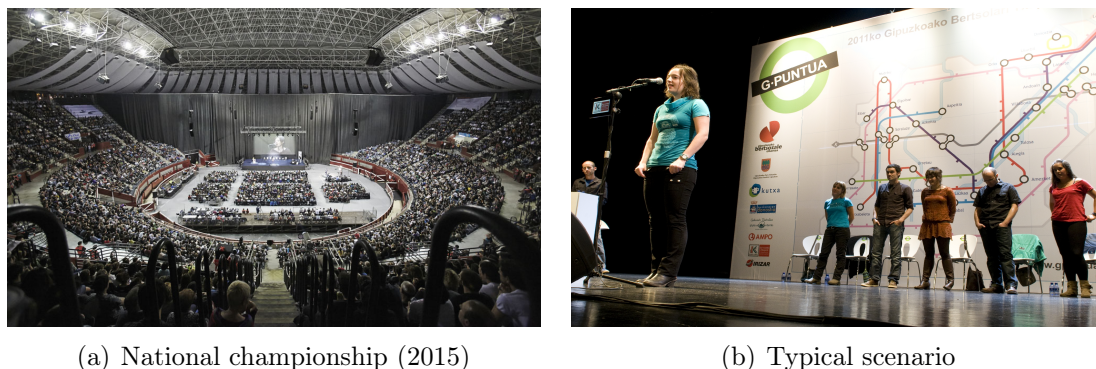


Figure 4.1: Description of the Bertsolaritza context

Xabier Amuriza, a famous verse-maker that modernised and contributed to the spread out of the *bertsolaritza* culture, defined *bertsolaritza* in a verse as:

<i>Neurriz eta errimaz</i>	<b>Through meter and rhyme</b>
<i>kantatzea hitza</i>	<b>to sing the word</b>
<i>horra hor zer kirol</i>	<b>that is what kind of sport</b>
<i>mota den bertsolaritza.</i>	<b>bertsolaritza is.</b>

Different poetry disciplines similar to *bertsolaritza* can be found around the world, such as Italian bards, Argentine payadors or Catalan glossators. However, the closest example is the American poetry slam, another oral poetry contest in which poets read or recite poems and are judged by selected members of the audience, and sometimes also by a panel of judges, like in *bertsolaritza* championships.

The art of composing extemporaneous verses requires a number of formal prerequisites that must be taken into account. Rhyme and meter are inseparable elements in improvised verse singing. A person able to construct and sing a *bertso* with the chosen meter and rhyme is considered as having the minimum skills required to be a *bertsolari*. But the true quality of the *bertso* does not only rely on those demanding technical requirements, the real value of the *bertso* resides on its dialectical, rhetorical and poetical value. Thus, a *bertsolari* must be able to express a variety of ideas and thoughts in an original way while dealing with the mentioned technical constraints. In this balance lies the magic of a *bertso*. Moreover, *bertsolaritza* belongs to oral poetry genre, which implies that a work has to be composed and performed at the moment, with no prior preparation. Performing in public is extremely important in such context, because the verse generation process is influenced by multiple factors perceived at each specific instant by the “actor”.

From the point of view of social robotics, we consider that *bertsolaritza* offers another sphere for improving social behaviours in robots. It can be an appropriate context to develop robot body language and robot communication capabilities for humanoid robots. Thus, we have defined an experimental goal to this research work:

design and develop a framework for a *bertsolari* robot, named Bertsobot. But the aim of this research work goes beyond to develop a robot capable of singing verses, we would like to take a step forward in the communication between humans and robots, designing and implementing the set of behaviours the robot needs in the stage to increase in one hand robot autonomy, and in the other hand, credibility and sociability of robots. In other words, we want our Bertsobot to behave social.

### 4.3 The ecological niche of Bertsobot

As mentioned before, we will focus on developing minstrel robots that sing improvised poetry to the public in Basque. If our robots are meant to participate in such events, they require certain capabilities that allow them act as human *bertsolari*-s.

The first step is to analyse the behaviour that troubadours show on stage. In *bertsolaritza* there are different type of public performances, such as tribute sessions in which one or several *bertsolaris* take part, free sessions between only two or three troubadours, or *bertso-saio*-s (sessions) guided by a presenter in which she/he proposes different exercises to the *bertsolari*-s. The dynamics of those performances differs from one to another, but here we will focus on the last one, that is the most typical scenario. Figure 4.1(b) gives a good overview about such scenario, composed by several chairs and microphones, some *bertsolari*-s, an emcee, and the audience.

The dynamics of a public performance guided by an emcee can be summarised in five main steps:

1. Wait sitting for its turn.
2. When it is its turn, approach the microphone.
3. Listen to the emcee and compose and sing the mandated set of verses to the public.
4. Observe and receive audience's feedback and react accordingly.
5. Go back to the sitting place.

These five steps describe the basic flow of an event, and therefore, the general behaviour that Bertsobot must show on stage, i.e. the ecological niche [123] of the robot. It is not a sequential process, each *bertsolari* can be called several times by the emcee, either alone or together with one or more fellows, and each time they can be mandated to sing several verses, i.e. steps 3 to 4 are looped. Moreover, on each step or phase of the performance the system must accomplish different tasks. For instance, in the second step the robot has to recognise its turn to sing, get up from the chair and find the microphone, move towards the microphone, listen to the exercise, and obtain the exercise requirements (topic or rhymes).

Having explained the general behaviour of the robot, the next step is to define and implement the behaviours that endow Bertsobot with the required capabilities to perform the actions associated to every phase of the performance.

## Chapter 5

# Describing the basic behaviours of Bertsobot

This chapter pretends mainly develop the basic capabilities that Bertsobot needs to be socially believable while acting as human troubadour. Those capabilities are going to be integrated in a robot framework that will be composed by several behavioural modules. From now on, the terms “framework” and “architecture” will appear several times throughout this report. We will use both interchangeably to refer to the implementation of the architecture that describes the structure and internal connections of the behavioural modules defined for our robots. There is no standard framework that defines which are the specific capabilities that a social robot must show while interacting with humans, but such capabilities indicate how effectively such agents interact. According to Dautenhahn [41] social interactive robots should exhibit the following characteristics: express and perceive emotions; communicate with high-level dialogue; learn models of other agents; establish and maintain social relationships; use natural cues (gaze, gestures, etc.); show distinctive personality and character; and learn social competencies.

While reviewing other existing robotic frameworks to take inspiration for our Bertsobot’s framework, we found several papers introducing human-robot interaction and social interaction-related frameworks that include some of the characteristics aforementioned. For example, Sarabia et al. [138] present a generic middleware for prediction and recognition of human actions and intentions, showing its application in a social context where the robot recognises and imitates human dancing movements. A framework composed by a bundle of ROS modules for HRI, named HRItk, is proposed by Lane et al. [96], which allows for simple interaction capabilities like speech recognition, natural language understanding, and basic gesture recognition as well as gaze tracking. Dias et al. [44] present a generic and flexible modular architecture for emotional agents with planning capabilities, designed to use emotions and personality to influence the agent’s behaviour. Jang et al. [82] propose a social HRI framework that provides two primary elements: cognitive capabilities for perceiving and interpreting social situations and planning socially appropriate actions, and high-level semantic interfaces for sensing and control capabilities. Finally, Fischer et al. [58] provide a software framework for the iCub

robot which integrates several components related to perception (object recognition, agent tracking, speech recognition, and touch detection), object manipulation (basic and complex motor actions), and social interaction (speech synthesis and joint attention).

As we have noticed throughout the review of the literature, the characteristics considered as essential are: verbal communication, non-verbal communication and perception of the environment. Moreover, we have also identified those related to perceiving and showing emotions, and the emotional state of the robot. In the following sections we will focus on defining the minimum skills or basic behaviours that constitutes the Bertrobot's architecture; the behaviours related to verbal communication, perception of the environment, and non-verbal communication capabilities will be introduced. Those basic behaviours will be explained without considering an important factor, the emotional state of the robot. How the robot perceives and shows emotions, and how the emotional state affects the behaviour of the robot will be explained later on in Chapters 7 and 8.

## 5.1 Verbal communication

One of the most natural way of interaction between humans is through speech. People use verbal communication to inform others of our needs, as well as to impart knowledge. If we intend human-robot social interactions, robots with speaking and understanding capabilities are essential. As previously discussed in Chapter 3, speech-based interaction means that the interlocutor and receiver use the same communication channel. Therefore, when designing a robot interface for verbal communication it is necessary to develop a TTS system for speaking capabilities and an ASR system for speech recognition.

Speech-based interaction is accomplished in two ways in our Bertrobot system: on the one hand, the system is able to maintain a dialogue with its interlocutor to receive instructions about the performance: when to start and when to finish, the theme and the metric to compose the poem, etc. On the other hand, oral communication is accomplished when the robot creates, under the given instructions, a new poem and sings it with a proper melody.

### 5.1.1 Speech-based dialogue

A Speech-based dialogue system is defined as computer system with which humans can interact through spoken natural language [63]. Its main purpose is to provide an interface between the user and the computer-based application that allows the interaction on a turn-by-turn basis. This definition covers a wide range of systems, ranging from simple question-answer systems to a more complex conversational systems. According to McTear [114], speech-based dialogue systems can be classified into three main types, depending on the methods used to control the dialogue with the user:

1. Finite state-based systems: the user follows a sequence of predefined steps or



states. At each dialogue state the system recognises specific words and phrases and produces actions based on the recognised response.

2. Frame-based systems: the dialogue flow is not predetermined but depends on the content of the user’s input and the information that the system has to elicit.
3. Agent-based systems: the dialogue model takes the preceding context into account, with the result that the dialogue evolves dynamically as a sequence of related steps that build on top of each other.

BertsoBot’s dialogue system allows the robot to maintain a simple conversation with the emcee. The dialogue is mainly guided by the emcee, who decides the turn to sing, the exercise and the topic or rhymes the troubadour has to employ to compose the verse, etc. The dialogue system proposed in this research work is composed by several modules that allow the robot to follow the dynamics of the dialogue with the emcee in a performance. Figure 5.1 depicts the architecture of our dialogue system that includes an “Automatic Speech Recogniser” (ASR), “Language Interpreter”, “Dialogue/Performance Manager”, “Response Selector”, “Text Generator”, “Speech Synthesiser” (TTS) and the “Singing Synthesizer” (TTSKantari).

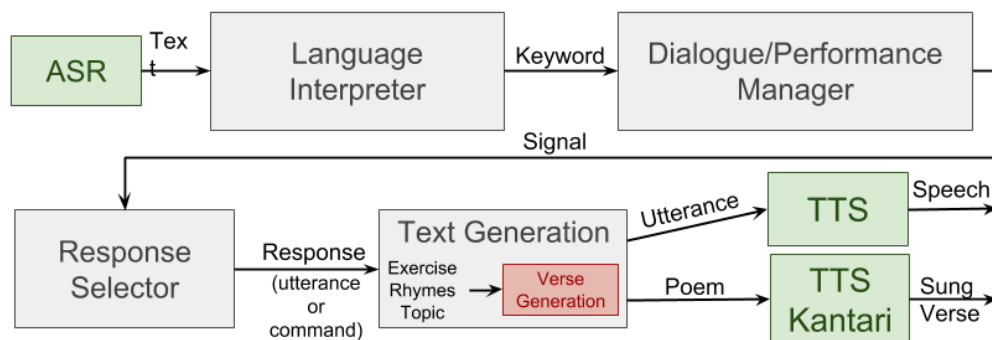


Figure 5.1: Description of the BertsoBot dialogue system

The “Automated Speech Recognition” (ASR) component converts the raw audio input into a sequence of words. Google Speech service<sup>1</sup> is used as ASR, which can be configured for many languages, including Basque. This is forwarded to a Language Interpreter module to extract the semantics of the utterance. The “Language Interpreter” module parses the input text and makes use of a database of keywords to identify user’s query. Then, the “Dialogue/Performance Manager” (PM) decides upon the action to take according to the employed dialogue strategy. The PM is the module that brings the coherence necessary to the system in order to follow the dynamics of the performance. Implemented as a finite state automaton, the PM defines the different phases of the event and controls the actions to be performed

<sup>1</sup><https://cloud.google.com/speech-to-text/>

at each state. The “Response Selector” selects the proper text output for the actual state. The output can be a predefined response according to the dialogue state (a set of utterances to receive information about the stage performance), or a command to compose a novel poem under the given constraints. The “Text Generator” module receives the input and generates the poem when it is commanded. Finally, the last step converts the text into audio. When the text output is an utterance it is passed to the text-to-speech engine (TTS) component to be synthesised. AhoTTS tool, a speech synthesizer for Basque language developed by AhoLab [76] is used for that purpose. Otherwise, when the output is a poem text, the last step consists of translating from text to a song that will be immediately performed by the robot.

To do so, poem’s metric is analysed and a melody is chosen randomly from an available database. The poem and the melody are sent to the TTSkantari singing synthesizer [1] which produces the audio file with the sung poem.

This is an illustrative example of user-robot dialogue in a stage performance:

HUMAN : Hello /robot/. Come to the microphone please. It’s your turn.

*(The robot stands up, reaches the microphone, and looks around until it finds the emcee. It gazes to him/her.)*

ROBOT : I am ready. Tell me, what exercise do I have to do?

HUMAN : Nothing is so beautiful as spring, that is the theme to compose a poem.

ROBOT : Sorry, but I did not understand what you said.

HUMAN : I will give you a theme and you have to compose a poem about it.

ROBOT : The exercise is theme-given. Great. What will be the theme of the poem?

HUMAN : The Spring. The theme to compose the poem is the spring.

*(The robot takes few seconds to create the poem and sings it)*

### 5.1.2 Verse generation

The core element of the “Text Generation” module is the “Verse Generation” system. Our approach implements the same strategy used by *bertsolari-s* for the creation of impromptu verses, and in a few seconds - less than a minute - assembles a new poem along the prescribed verse-form. The proposed system receives as input the type of exercise and the topic or four rhymes (depending on the exercise) and tries to give as output a novel poem that: (1) satisfies the formal constraints of rhyme and metric, and (2) shows coherent content related to the given topic.

The poetry generation strategy employed is a corpus-based method [10] and the overall semantic relationship has been implemented with an Latent Semantic Analysis (LSA) model [42][8]. The verse generation procedure relies in the extraction of sentences from corpora and combining them (under rhyme and metric constraints) to form the final poem. The LSA model assures the internal coherence between poem lines and the overall coherence with respect to the theme.

The two text generation methods are:

- Sentence retrieval: The basics of this method is to extract from the corpus

sentences which meet rhyme and metric constraints.

- N-gram probabilistic model: Starting from the rhyming word, the verse is built backwards using the selected N-gram model; extending at each step the sequence of words with new ones that have a non-zero probability of appearing after the last word.

Figure 5.2 depicts the verse generation system. A detailed explanation of the verse generation process and the tools developed for that purpose can be found in Astigarraga’s research work [6].

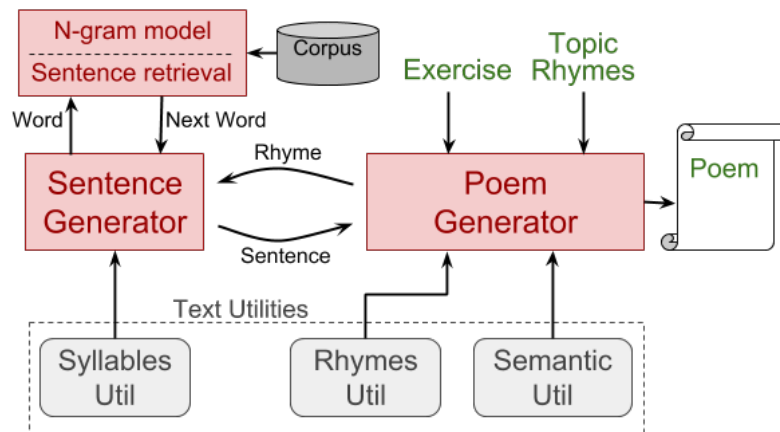


Figure 5.2: Verse generation system

It must be noted that the goal of the verse-makers is not only to convey a message in the form of a poem but also to respond to an affective target and/or to create an affective response in the audience. How audience reactions are processed and coded and how are used to adapt the emotional behaviour of the robot is addressed in Section 7.2.

## 5.2 Perception of the environment

Embodied cognition establishes that cognition depends upon experiences that come from having a body and thus, feedback between agents and the world is essential to develop cognitive capabilities [37]. In order to interact with its environment, a social robot must be able to perceive it. Perception is fundamental for the robot to detect changes and react to the stimuli.

BertsoBot is able to interact with the environment in different ways, it can identify and recognise the key elements of the scenario, locate voice sources, and orient the gaze to the interlocutors’ faces. BertsoBot’s perception is composed by two main modules: “Key Objects Perception” and “People Perception”. The former allows the robot to find and track the key elements on stage, and the latter permits the robot to perceive people and direct its gaze towards the interlocutor.

### 5.2.1 Perception of the key elements

The robot pays attention to different elements at different steps. The robot can be requested to reach the microphone to start its singing turn or it may need to go to rest to its chair. For the time being, those elements, as well as being adapted to the robot’s morphology, they have labels to make it easier the identification and recognition processes. They all have colour tags that make them distinguishable; chairs have been painted with different colours and, similarly, the microphone has a blue tag on its base. Every key object has a QR code to make it recognisable. Our approach for object identification combines a colour tracking algorithm with QR detection. The colour tracking procedure enhanced with a Kalman filter<sup>2</sup> is used to produce a more robust behaviour against illumination conditions and balancing produced during walking. The QR detection is done by using the ZBar bar code reader library<sup>3</sup> that gives the text related to the QR codes detected in the source image. No location information in form of odometry or frame of reference is used because the location of those elements with respect to the robots varies depending on the scenario.

### 5.2.2 People perception

A natural reaction when we want to interact with someone is to direct our gaze towards the interested agent. The gaze feeds the communication, and conveys interest or attention to the interlocutor. It requires positioning the robot to make the most out of its sensors and to let the human talker know what the robot is actually paying attention to. Research in social interactions has investigated that displaying appropriate human-like gaze behaviour improves people’s perceptions of the conversational effectiveness of humanoid robots [4][163].

Spontaneity during verbal communication involves two main behaviours, face and sound localisation, and BertsoBot’s “People Perception” module combines both to perceive and track the presenter.

Face localisation is done applying OpenCV’s Haar feature-based cascade classifiers [166] to the images taken by the upper camera on the robot’s head. Once the face is detected within an image, the centre of the face ( $CF_x, CF_y$ ) in the image is obtained, and the head joint angles (pitch and yaw) to track the face, with respect to the centre of the image, are calculated as shown in equations 5.1 and 5.2.

$$H_{pitch} = (fov_{vertical}/img\_size) * (\frac{num\_rows}{2} - CF_x) \quad (5.1)$$

$$H_{yaw} = (fov_{horizontal}/img\_size) * (\frac{num\_cols}{2} - CF_y) \quad (5.2)$$

Sound localisation allows a robot to identify the direction of sound, and it is done using Aldebaran’s “ALSOUNDetection” algorithm based on Time Difference of Arrival (TDOA) approach [22]. The sound wave emitted by a source is received at

<sup>2</sup>[https://docs.opencv.org/trunk/d1/da2/kalman\\_8cpp-example.html](https://docs.opencv.org/trunk/d1/da2/kalman_8cpp-example.html)

<sup>3</sup><http://zbar.sourceforge.net/>

slightly different times on each of the robot's four microphones, from the closest to the farthest. These differences are related to the current location of the emitting source. By using this relationship, the robot is able to retrieve the direction of the emitting source (azimuth and elevation angles) from the TDOAs measured on the different microphone pairs. In this case, pitch and yaw angles to track the sound with the robot's head are obtained as shown in equations 5.3 and 5.4:

$$H_{pitch} = head_z + elevation \quad (5.3)$$

$$H_{yaw} = head_y + azimuth \quad (5.4)$$

### 5.3 Non-Verbal communication: Body language

Verbal communication and non-verbal signs come together in humans; verbal communication is the most natural communication way that we use for social interaction, but it is the non-verbal communication what really helps us to understand sociability [90]. Much of non-verbal communication is unintentional, people are not even aware that they are sending messages through body gestures and postures, facial expressions, eye contact, and head movements.

Social robots must be expressive in a human-like way in order to be socially accepted. In order to achieve the most basic degree of naturalness any humanoid robot must be endowed with some of the non-verbal signs mentioned above. For the moment we will focus on how to enhance the expression capabilities of our BertsoBot through body gestures.

L'homme and Marsella [100] discuss body expression in terms of postures, movements and gestures. Gestures, defined as movements that convey information intentionally or not, are categorised as emblems, illustrators and adaptors. Emblems are gestures deliberately performed by the speaker that convey meaning by themselves and are again culture dependent. Illustrators are gestures accompanying speech, that may (emblems, deictic, iconic and metaphoric) or may not (beats) be related to the semantics of the speech [113]. Lastly, adaptors or manipulators belong to the gesture class that does not aid in understanding what is being said, such as ticks or restless movements.

When *bertsolari*-s are on stage they are continuously conveying information, through facial expressions, body gestures, or head movements about their mental state. The body expression of the troubadours changes depending on its mental state, that varies throughout the performance. After identifying the main different states of the global behaviour, a gesture library composed by a set of adaptors and a set of illustrators has been defined to mimic verse-maker's expressiveness on stage: the former set consists of waiting gestures, thinking gestures and singing gestures, that refer to a different mental state of the troubadour, while the latter set includes those gestures that accompany the speech (specifically, beats).

### 5.3.1 Waiting gestures

Humans are not designed to be motionless while awake, and so, it is not appropriate to have a robot sat inert or stood up still on stage. Humans stretch or cross their legs, drink water or move their head to change their gaze while sitting. Without doubt, our robots' movements are very limited in that position, and most of the mentioned moves cannot be replicated. But they can change their arms' position and make movements with their heads. Figure 5.3 shows an example of the gestures that the robot performs while remain sat on their chair waiting their turn to sing.

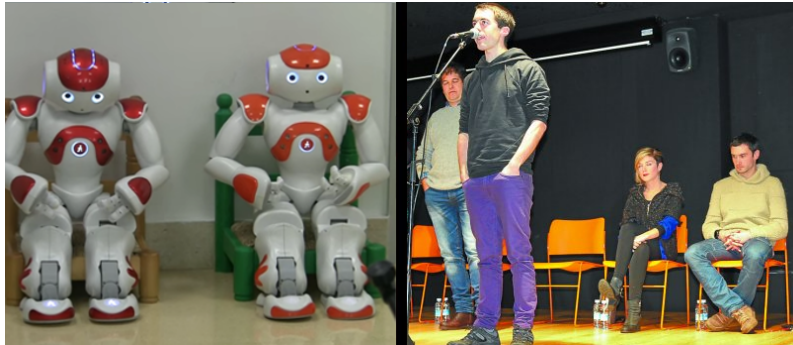


Figure 5.3: Example of a waiting gesture

### 5.3.2 Thinking gestures

Those gestures that troubadours unconsciously make while standing up in front of the microphone and thinking the verse (see Figure 5.4). They are movements to unstress, to relax tension, such as putting one's hands behind one's back, swinging the hip, scratching one's head, etc. There is one extremely important gesture while thinking: reaching and maintaining a neutral pose. The robot needs to move, needs to reproduce some gestures but it cannot be continuously gesturing like a puppet; improvising a verse is a very hard mental process that requires extreme concentration and that is reflected in the body language of the performers.

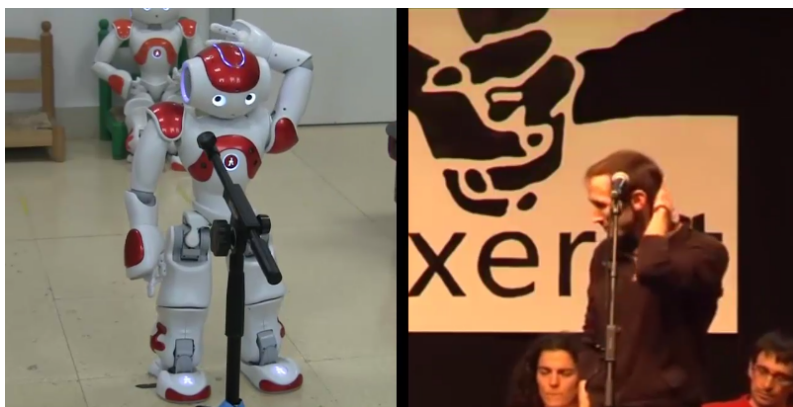


Figure 5.4: Example of a thinking gesture

### 5.3.3 Singing gestures

Just after the improvisation process finishes and before the *bertsolari* starts singing, they need to accommodate the body and/or clear the throat, look around and probably stare off into space, above the public (see Figure 5.5). Oddly, and probably due to the extreme concentration effort that must be maintained, the verse-maker stands still while singing. Of course, not everyone maintains the same pose, sometimes they keep their hands in their pockets, or behind their back, or just have their arms hanging down, but that pose does not vary significantly from one *bertsolari* to another. Thus, no gesture is reproduced while singing.

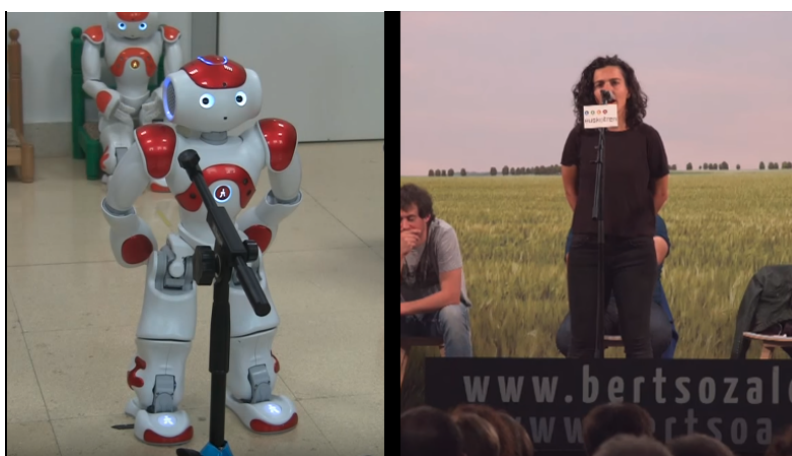


Figure 5.5: Example of a singing gesture

### 5.3.4 Talking gestures

Humans don't stay still while talking, we naturally gesticulate by moving the hands. Our robots also accompany their speech by moving their arms. The gestures that the robot perform while talking are not associated with particular meanings, they occur with the rhythm of the speech. The first approach we used to generate talking gestures was simple, the sequence of the predefined gestures to be executed was randomly selected and the number of gestures was chosen according to the duration of the speech. Those gestures were obtained from NAO and Pepper's animation library. Another two approaches that enhance the behaviour of the robot while talking will be described later in Chapters 7 and 8.

On the other hand, as well as humans nod when they are talking with other people, our robots also move their head up and down to make it know that they has understood something. It does not mean that they know what has been said, but it makes the interlocutor realise that the robot has successfully processed the captured audio.

## 5.4 Contributions and conclusions

The main contribution related to this chapter is a set of behaviours that allows BertsoBot to perform the following tasks:

1. Communicate in natural way with the emcee and with the other contestants.
2. Identify some environmental key objects.
3. Compose and sing verses based on demanded technical requirements.
4. Show human troubadours-like body expression on stage.

Those behaviours are the principal components of the BertsoBot’s framework, which has been developed using ROS. The framework has been defined as a ROS based control architecture that endows the robots with some of the social capabilities that *bertsolari-s* possess, which here have been considered as the basic ones. The control architecture is composed by different behaviours or modules that make the robot act in a consistent manner and resemble to a real *bertsolari*. Those ROS modules are activated by different stimuli (speech orders, object detection, etc.) and depending on the state of the performance the robot executes the corresponding tasks. Different behaviours can be activated and combined at each state.

Figure 5.6 shows the state of the BertsoBot’s architecture at this point, including the behaviours that allow the robot to perform 1-4 tasks.

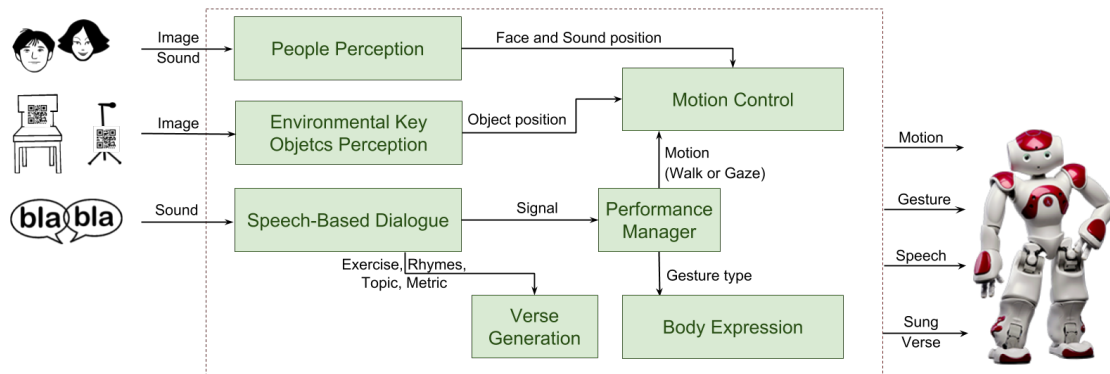


Figure 5.6: BertsoBot’s architecture with basic behaviours

Summarising Figure 5.6, the “Performance Manager” is the behaviour that brings the coherence necessary to the system in order to follow the dynamics of a performance. The “People Perception” as well as the “Speech-Based Dialogue” behaviours allow the interaction with the emcee, while “Environmental Key Objects Perception” provides the robot with necessary skills to interact with environmental key objects. These interactions, usually executed as motion actions (head or walking movements), are managed by the “Motion Control” behaviour. The verse is



composed and translated to a sung verse by the “Verse Generation” behaviour, and the robot body expression is managed by the “Body Expression” behaviour.

The set of basic behaviours defined throughout this chapter has been used and tested in several public performances. In April 2014 was the first time that NAO acted as verse-maker. It was in an event, called “Badu, Bada”, to which we were invited to give a talk about *bertsolaritza* and robots. NAO robot showed his verse improvisation and verbal communication capabilities, and it only gesticulates while thinking the verse. Later, in December of that same year, we also participate in “ScienceClub 2014”, an event that aim to disclose science and technologies to the society. A dialogue with NAO of approximately 10 minutes was presented, showing the same capabilities demonstrated in the previous event (see video<sup>4</sup>). We were also invited the following year to the “ScienceClub 2015” event. This time all gestures repertoire mentioned in section 5.3 were integrated and chatting abilities were shown. The key element recognition was tested together with the face and sound localisation behaviours (see video<sup>5</sup>).

The developed work resulted on the following publications:

- **Singing minstrel robots, a means for improving social behaviors.** I. Rodriguez, A. Astigarraga, T. Ruiz and E. Lazkano. 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, May 2016, pp. 2902-2907.
- **BertsoBot: Towards a Framework for Socially Interacting Robots.** A. Astigarraga, I. Rodriguez, T. Ruiz, E. Lazkano. Spanish Robotics Conference (JNR), Valencia, Spain, June 2017, pp. 117-122.

To conclude, the identified building blocks have shown to be enough for an effective behaviour. The developed architecture is able to produce the desired robot behaviour. The main limitation of the system at this point relies on the “Performance Manager”. It has been defined as a finite state automaton that operates in a open-loop, which requires the emcee to act as a sequencer, following successively all the phases/states of the performance. The initial setup of the system should always be the same (the robot started sat on the chair). Any discordance between the real state and the sequence of actions to be executed produces an undesirable global behaviour and requires the intervention of the operator and the interruption of the performance. In other words, the systems is fragile. This lack of autonomy is later on fixed by adding some level of self-awareness to the robot (see Chapter 6).

The BertsoBot’s gestures library has been defined using *Choregraphe*, a multi-platform desktop application developed by Softbank for creating applications and animations for NAO and Pepper humanoid robots. Except the talking gestures, that most of them have been collected from NAOqi’s animation library<sup>6</sup>. Each of the gestures of the library has been recorded with a predefined motion timing. In

---

<sup>4</sup><https://www.youtube.com/watch?v=HKxe40-Qi6w>

<sup>5</sup><https://www.youtube.com/watch?v=IMMXHWP2mZA>

<sup>6</sup>[http://doc.aldebaran.com/2-1/naoqi/audio/alanimatedspeech\\_advanced.html](http://doc.aldebaran.com/2-1/naoqi/audio/alanimatedspeech_advanced.html)

order to maintain public attention and interest, gestures cannot be predictable. Our approach to enhance the spontaneity of the robot consists of randomly select all, the number of gestures (between a delimited interval), the gesture set and the order in which they must be reproduced, at each state as the performance progresses. Although this solution seems a little naive, it has shown to be effective to increase the naturalness of the robot, from the perspective of the observer and thus, the empathy with the robot. This aspect will be improved by means of generative models (see Chapter 8).

## Chapter 6

# Improving autonomy of social robots by self-awareness

As social robots are increasingly endowed with human natures (e.g., voice, appearance, and motion) and applied in different contexts (e.g., education, care, entertainment), it is desired that robots are able to accomplish the tasks for which they have been designed (e.g., teaching children, supporting elderly, singing verses) by themselves and without surveillance. This idea implies a certain level of autonomy.

According to the Merriam-Webster<sup>1</sup> dictionary, autonomy is the quality or state of being self-governing. However, the concept of autonomy in robots goes further and comprises many qualities, such as long term functioning, adaptability, learning capabilities, operation with little human intervention, self detection of errors, etc.

There is agreement in the robotics community that autonomy is not a yes or no property; the degree of autonomy of a robot is a characteristic that ranges from no autonomy to high autonomy. Several definitions can be found in the literature about robot autonomy. Thrun [155] defines autonomy as the ability of a robot to accommodate variations in its environment. Patrick Rau et al. [130] define it as the degree to what a robot can act on its own accord. Rather differently, Bekey [20] refers to autonomy as a system capable of operating in the real-world environment without any form of external control for extended periods of time. However, related to social robotics, the autonomy depends on its social roles, capabilities and the requirements expected from both itself and the other agents in its social environment [47].

Concerning HRI, a few attempts can be found in the literature about a taxonomy for measuring the degree of autonomy of robots. Yanco and Drury [174], identify several categories, such as task type, task criticality, robot morphology, and interaction roles, and define a taxonomy based on those categories. The autonomy level (A) of the robot and the amount of operator intervention required (I) are also considered. On the other hand, the autonomy level then measures the percentage of time that the robot is carrying out its task on its own; the amount of intervention required measures the percentage of time that a human operator must be controlling the robot. These two measures sum to 100%. Teleoperated robots are in the lowest

---

<sup>1</sup><https://www.merriam-webster.com>

level (A=0% and I=100%), and at the other end of the spectrum are robots with full autonomy like guided tour robots and delivery robots (A=near 100% and I=near 0%). In between these two points are shared control robots, those that are able to do some part of the task on their own but need the human operator help to perform other part of the task (A=75% and I=25%).

Alternatively, the LORA (Level of Robot Autonomy) framework [19] proposes a guideline to categorise the robot autonomy level in a 10-point qualitative taxonomy considering HRI variables such as trust and acceptance, reliability, and situation awareness. Again, the autonomy level is associated with the need for human intervention. Thus, robots that need to be operated by humans (teleoperated robots) to perform well have lower autonomy, and robots able to sense-plan-act with minimal human input are categorised as highly autonomous.

Some service robots are able to exhibit increasing levels of autonomy, capable of surviving and performing useful tasks (such as surgery or vacuum cleaning) in real environments for extended periods, but most of them performs under highly structured situations. However, at the present time, social robots are not fully autonomous. Although some robots in real life have impressive human-like appearance, none of them has a level of autonomy that comes close to that of humans. To increase the autonomy of social robots they should meet other certain capabilities that require awareness of their body. For instance being able to decide their own actions (walking, talking, gesticulating, etc.).

In the following sections we will discuss about self-awareness in robots, and how being aware of one's body can help to determine the next actions, and therefore, to improve the autonomy level.

## 6.1 Self-body awareness

According to Lagercrantz et al. [95] a simple definition of consciousness is the sensory awareness of the body, the self, and the world. The first thing children perceive is their own body, which serves as a means of interaction with others and the environment. Thanks to her/his body, the child experiences different sensations, mobilises and learns [167]. Recognising oneself, although it seems easy, at least requires to be conscious of one's body and one's actions. The five senses (sight, hearing, smell, touch and taste) are the traditionally recognised methods of perception responsible for our interaction with the external world. But additionally to the exteroceptive sensors, we have senses that are responsible for our internal functioning. That fundamental internal sensory system, called proprioception, provides feedback about the status of the body internally without the aid of vision. Thanks to muscle spindles, which detect changes in the muscle and signal the angle of related joints, we get information about limb positions, and we realise our body's position.

Just as humans show consciousness of their body, social robots, in the way to be truly autonomous, should also be able to recognise their own configuration. Several robotic systems can be considered as self-aware systems to some degree, being able to recognise themselves in the mirror [73], or being aware of their motion [115]. In

[13] authors show a bio-inspired somatosensory system for a humanoid robot based on a set of soft sensors that allows the robot to perceive and interpret physical sensations. Finally, Zeng et al [175] propose a brain-inspired bodily self model that permits a NAO humanoid robot to recognise its own body in real world and in a mirror.

As mentioned before in Section 5.4, the main limitation of the Bertsobot system relies in the “Performance Manager”. The emcee must act as a sequencer and the robot must execute successively given orders, starting the sequence always from sat on the chair position. The robot does not have any mechanism to detect any discordance between the real state of the robot and the sequence of actions to be executed. Consequently, undesirable robot behaviour can occur when the robot is not able to fulfil the action requested. For instance, when the robot is called to approach the microphone it has to perform the following sequence of actions: stand up from the chair, find the microphone, and move towards the microphone. If it fails in any of those actions the interruption of the performance is required and the system must be restarted.

In order to overcome this limitation Bertsobot must be able to “decide” its own actions, of course limited to the orders given by the emcee. To do so, it requires an action selection mechanism that will plan the sequence of actions to be executed depending on the current posture of the robot and the state of the performance. We propose a body self-awareness behaviour for the Bertsobot based on robot posture recognition. This approach is described in the following sections.

### 6.1.1 Posture recognition

Humanoid bodies with human-like gesticulation capabilities enhance the body expressiveness of the robot, and hence, the effectiveness of the interaction between humans and robots. A social robot must also be aware of what happens in its body before performing any action. Feasible gestures, movements and actions depend on the current body posture. Multiple works can be found related to human posture recognition: approaches based on depth images [143][170] or approaches that rely on skeleton information [131][35], to mention some.

All works mentioned above, both self-aware systems and human posture recognition related works, are based on external perception. We tackled the problem in a rather different way, taking into account only internal sensory receptors; the body posture recognition system proposed here allows Bertsobot to know self-body posture without the aid of visual information. It must be noted that this part of the work was mainly developed for NAO. According to the flow of a performance two are the principal postures the robot can show: sat on a chair and stood up (see Section 4.3). But the robot can often initialise in a typical comfort pose like crouched or sat on the floor. Thereby, four different body postures have been defined as body states to be recognised: sat down on the floor, sat on a chair, stood up and crouched.

The taken approach for posture recognition relies on a skeleton based approach that uses proprioceptive sensors. Every posture of the robot is defined as a vector including the positions (x, y, z) and the Euler angles (roll, pitch, yaw) of each

joint of the robot. Those values are obtained from the robot’s coordinate system. The problem of recognising the posture is one of classification: different postures are associated to different vectors and given a vector, the robot position has to be inferred.

This problem lies in the field of supervised learning, where a model is built from a training set of known instances to predict the correct class of new presented instances. Several classification methods (decision trees, Naive Bayes, K-Nearest neighbours and Support Vector Machines) have been considered to build our model for posture recognition. Among the mentioned classifiers C4.5 [128] and 3-NN [2] slightly outperformed. We need fast response from the classification system because the robot is moving permanently and thus, the non lazy nature of the classification tree made it more suitable for our needs.

Once the model is built, the posture recognition process is quite simple. It can be summarised into two main steps as shown in Figure 6.1. In the initial step, required features are obtained with respect to the robot’s main reference system. And then, the acquired decision tree model is applied to the extracted data, which returns as output the predicted posture of the robot.

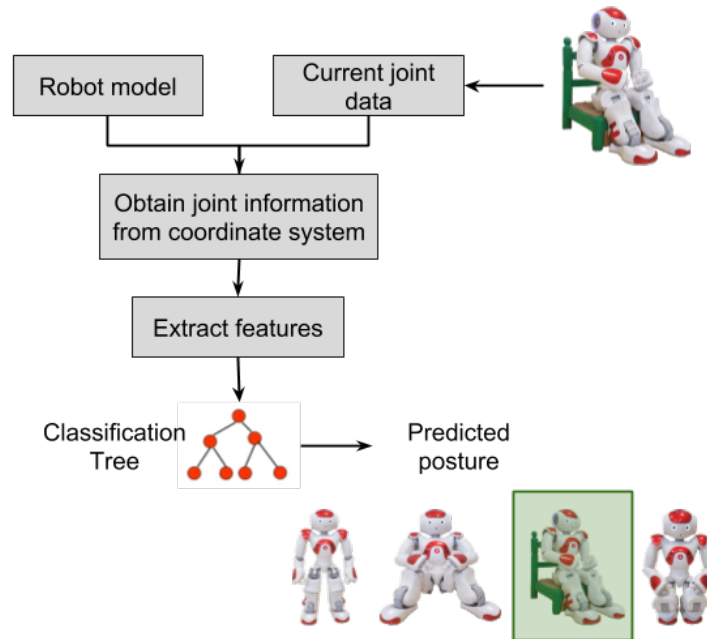


Figure 6.1: Posture selection mechanism

### 6.1.2 Action selection

According to Mataric an autonomous robot should be able to take its own decisions in order to fulfil its goals [110]. To this end, it must know the tasks or the actions to execute in each situation. The autonomy is then related to the selection of actions during the robot’s “life”. Castro et al. [33] state that the behaviour of the robot

differs according to the policy that determines the next action that will be executed at each moment. This policy can be acquired in two different manners:

1. The policy is assigned and the robot follows this pre-designed policy.
2. The robot learns the best policy according to certain requisites.

Regarding Bertsobot’s decision making capability, the action selection procedure employed up to now was quite simple: the robot executed the action or the sequence of actions associated to the order given by the emcee. Those actions were hardcoded and executed sequentially without considering whether it was feasible or not for the robot, due to its current body posture. Therefore, the robot’s autonomy to determine the action to perform was minimal.

Aiming to enhance the autonomy level of Bertsobot, we have developed an action selection mechanism that determines robot’s actions algorithmically. Its main task consists of translating the order given by the emcee into an action or sequence of actions that are selected depending on the current posture of the robot. In this way, the robot has no longer to start always the performance in a unique valid position, and the presenter has flexibility to change the flow of the performance on the fly. Moreover, the action verification process allows the system to notify the presenter about some failures when executing actions, such as it has not been able to get up from the chair or to find the microphone.

## 6.2 Contributions and conclusions

The main contribution of this chapter is the self-awareness behaviour (called “Body Self-Awareness”) developed for NAO robot. This behaviour is composed by two main modules, the action selection and the body posture recognition modules. Figure 6.2 depicts Bertsobot’s architecture after introducing that behaviour. Improvements have been highlighted in the figure with red colour.

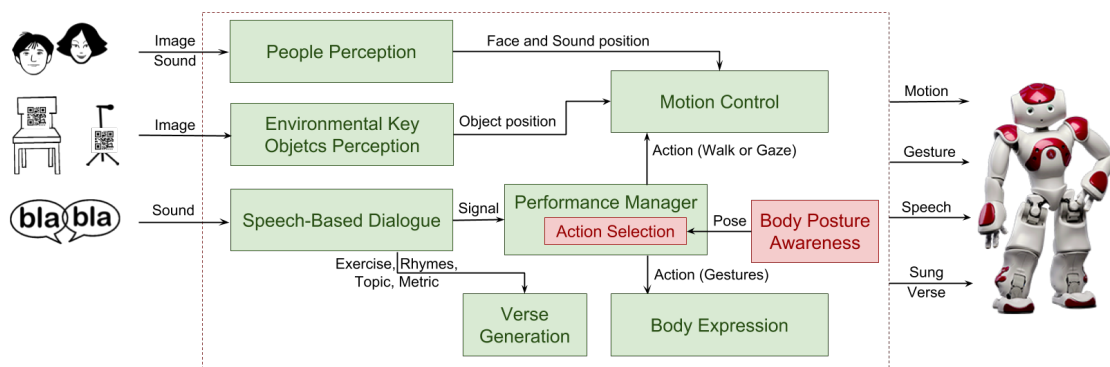


Figure 6.2: Bertsobot’s architecture including self-awareness behaviour

Now, “Performance Manager” includes an action selection mechanism (“Action Selection”) that together with the posture recognition system (“Body Posture Awareness”) allow the robot to take decisions in a more autonomous manner. The “Action Selection” module receives the order and translates it into an action or sequence of actions to perform according to the current posture of the robot, which is given by the “Body Posture Awareness”. Each action emerges in a gesture or a movement, and must end in a specific posture. After the execution of the action, “Action Selection” verifies that the gesture/movement ends successfully and it checks whether the actual posture of the robot corresponds to the expected one. It also controls the postures in which walking is possible. For instance, if the robot needs to reach the microphone, it needs to stand up before it starts to walk. This behaviour verifies this condition and commands the proper movements (depending upon the current posture) to the “Motion Control” module to perform desired movement, or to the “Body Expression” module to execute the desired gesture.

Therefore, the improvements introduced in the overall system are three-fold:

1. The robot gestures always match the gestures allowed by its current body position and thus, no weird behaviour occurs.
2. There is no need to execute the same sequence of actions during the performance. The emcee is free to change the flow in real time and the initial setup is unnecessary. The robot adapts the actions to perform accordingly to its current state independently of the initial posture.
3. More complex actions can be defined that comprise several sub-goals. The emcee does not need to worry about each next step anymore.

We empirically show the improvements introduced in the system by incorporating the “Body Self-Awareness” module by making and recording two experiments:

- Experiment 1: the robot receives a “stand up” order from every different pose. The video<sup>2</sup> shows how the robot adjusts the movement to be executed depending on its current body posture to obey the order given by the human. NAO nods to confirm it has understood the command and when it reaches the desirable final posture it says so.
- Experiment 2: the *bertsolari-s* are not always called to approach the microphone directly while sitting on the chair, they can be required to first stand up and listen to the exercise, before they approach the microphone to sing. In this second video<sup>3</sup> the robot is told to reach the microphone, again from different initial postures. The complexity of this order (as that of sitting on the chair) is higher because the reference element does not need to be in the robot’s current field of view.

<sup>2</sup><https://www.youtube.com/watch?v=x88fK8luYMc>

<sup>3</sup><https://www.youtube.com/watch?v=0uax6qilK30>



Bertsobot's body self-awareness behaviour is deeper explained in the following publications:

- **Body Self-Awareness for Social Robots.** I. Rodriguez, A. Astigarraga, T. Ruiz and E. Lazkano. 2017 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Prague, Czech Republic, May 2017, pp. 2902-2907.
- **On how self-body awareness improves autonomy in social robots.** I. Rodriguez, J.M. Martínez-Otzeta, E. Lazkano, T. Ruiz and B.Sierra. 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), Macau, China, December 2017, pp. 1668-1693.

Finally, the work presented here is related to the long-term goal of building a social robot with consciousness, especially with self-awareness. The “Body Self-Awareness” behaviour proposed here permits to derive robot's action algorithmically instead of being prescribed by a human, adapting those actions/gestures to whatever its body posture is. As mentioned in the introduction of this chapter, autonomy is a gradual property not easy to be measured neither quantitatively nor qualitatively. But the newly integrated module gives the system robustness and at the same time, makes it more flexible. It contributes to ease the work of the emcee and basically, it increases the system's autonomy. We believe that this module entails a big improvement in the overall behaviour.



# Chapter 7

## Making our robots affective

Emotions are an important part of multimodal communication, and strongly affect social interactions. When we interact with other people, we transmit our emotional state, giving clues about how we are feeling and, at the same time, the emotional expressions of those around us give us a glimpse into their inner mental state. Emotional expressions can occur with or without self-awareness during the interaction, and can be manifested in different ways, such as facial movements, body postures, gestures, etc. Those expressions are also often used to support the verbal communication, accompanying the speech and conveying in turn information about the emotions and thoughts of the sender.

Designing robots able to perceive and show some kind of emotions can make social interactions between humans and robots more effective and natural [80]. Robots with expressive characteristics have benefits for people in two ways: by communicating emotions to humans and by influencing humans' behaviour. Systems that can both be influenced by and influence users exhibit an affective loop experience [77]. The affective loop is the interactive process in which the user of the system first expresses his/her emotions through some physical interaction involving his/her body, and the system responds by generating emotional expression. This expression in turn makes the user respond and feel more and more involved with the system (see Figure 7.1). Two requirements are needed to establish this affective loop between users and robots [120]. On the one hand, robots require an emotion perception system that recognises, among other states, whether the user is experiencing positive or negative feelings. On the other hand, a reasoning and response selection/generation mechanism is needed that chooses the emotional response to display at the cognitive level.

In this chapter we will describe the two emotional modules developed that will make our social robots affective. The first one is related to the emotional response that the robot must show after singing a verse. If the troubadour performance is to be perceived as credible, lively and creative, public reaction must be sensed somehow by the robot and its behaviour must reflect the noticed sensations, either integrating them in the next sung verse like real troubadours do or showing a proper body language. The second one refers to the expression capability of the robot during talking. In order to improve the naturalness of the robot during the interaction

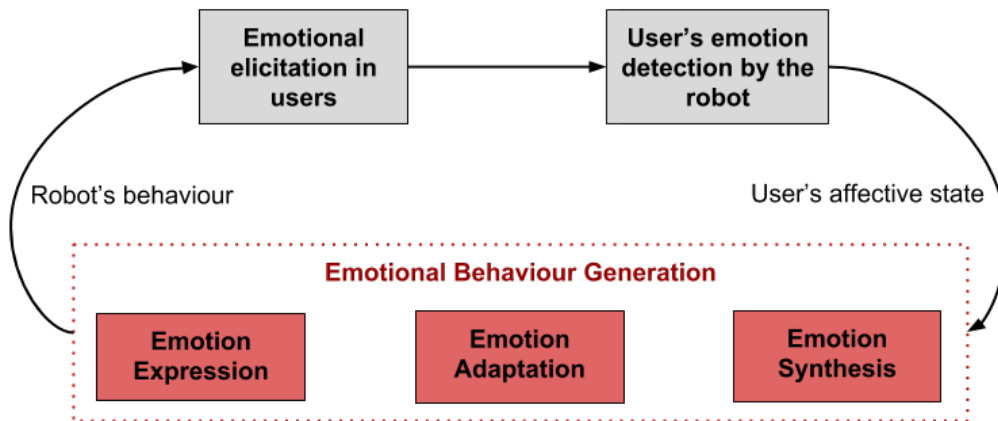


Figure 7.1: Graphical description of the affective loop process

with the user, the robot must be able to talk and gesticulate showing appropriate emotions that fit the emotional content of the spoken text. Thus, the affective talking behaviour allows us to give a step forward and go beyond the BertsoBot, by integrating expression into one of the basic skills social robots should show: verbal communication.

## 7.1 Emotion theories

When it comes to developing an emotional system for a robot, two questions must be answered: what emotions are going to be considered? and how must those emotions be represented?

The theory of basic emotions states that emotions can be divided into discrete and independent categories [49]. Paul Ekman identified six basic emotions (anger, disgust, fear, happiness, sadness and surprise). Alternatively, dimensional affective models regard affective experiences as a continuum of highly interrelated and ambiguous states. Emotions are described as linear combinations of *Valence-Arousal-Dominance* (VAD). *Valence* defines how positive or negative the stimulus is, *Arousal* specifies the level of energy and *Dominance* defines how approachable the stimulus is. These models allow for a wider range of emotions [135].

We must note that at this point we do not have defined a global emotional state that drives the behaviour of the robot combining several inputs. The aforementioned emotional behaviours work independently, and each behaviour employs a different strategy to select the emotion to be displayed by the robot.

## 7.2 Understanding and reacting to audience feedback

Audience plays an important role in live performances, specially in *bertsolaritza*. The crowd shows how pleasant the verses have been, usually clapping as well as laughing when they have found it amazing. *Bertsolari*-s create the verses on the spot, based on current perceptions, and the feedback of the audience strongly affects in such composition process.

Developing an approach to react to the audience's feedback covers multiple fields, such as applause detection, classification, selection of the robot's appropriate reaction in the context of the performance and generation of messages with predetermined sentiment.

The goal of this work is to move on to a closed-loop form of the robot performance where the robot perceives audience's feedback measuring the intensity and duration of the audience's applauses, and in response the robot changes its mood and tries to better please the public. The change in its mood is reflected on the one hand, in its body expression by reacting through subtle gestures, and on the other hand, in the composition of next verse by adapting the sentiment of the message.

### 7.2.1 Obtaining emotional feedback from audience applause

The problem of content-based audio classification and segmentation has been studied intensively outside the field of robotics and some work has specifically focused on applause. Cai et al. [30] have successfully used Mel-Frequency Cepstral Coefficients (MFCC) and a set of low-level features such as sub-band energies to find significant audience reactions including applause and laughter.

Few works have been done when it comes to observing robot induced audience expressions. Knight et al. [91] have developed a stand-up comedian robot that varies joke selection depending on pre-communicated visual feedback and noise level. Another performance robot by Katevas et al. [88] similarly features joke-telling. It incorporates visual emotion recognition and detecting the noise levels to delay the performing of the comedy script. Audience feedback is partly elicited by the robot itself leaving the spectators in a natural comedy setup without human interference.

Our approach for detecting the emotional state of the public uses audience applause as feedback to the robot system. Applauses are captured and translated into a response from the public by means of energy (E) and duration (d) of the applause. The addressed strategy can be split up into a straight-forward work-flow (see Figure 7.2). In the initial step, audio processing and machine learning techniques prepare the input audio stream by first chunking it, and then classifying each chunk as being applause or not. Several supervised classification algorithms were tested, and C4.5 decision tree was selected as the audio classifier. Next, the incoming stream of classified chunks is segmented into sections of consecutive applauses, leading to a small descriptor ([E, d]) for every evaluated applause. Based on all previous applauses of the event, the most recent one can subsequently be classified. The applauses are

coarsely categorised as belonging to one of the following classes: Negative, Neutral, Positive and Very Positive.

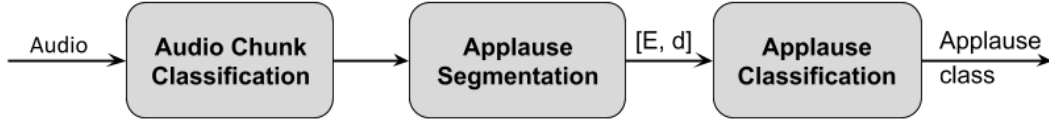


Figure 7.2: Approach work-flow for applause classification

As all events are different, due to enormous variety in audience sizes, acoustic perception and emotional state of the audience, it is difficult to compare them. Thus, in our approach for applause classification we decided to use unsupervised online learning techniques. We use k-means to do clustering with a variable number of classes. The first applause is always classified as Neutral, after that, the number of available classes is increased.

### 7.2.2 Emotional response through gestures

Once BertsoBot is able to perceive the audience’s emotional reaction, the system needs to choose the proper emotional response to show, it must be one that mirrors the audience’s acceptance of the sung verse. This process requires a reasoning step and a displaying step.

Our approach to select the robot’s emotional state is based on a direct translation from the class obtained in the applause classification process into an emotion. Table 7.1 shows the correspondence between the applause class and emotion.

Applause class	Robot emotion
Negative	Sadness
Neutral	Calm, serene
Positive	Pride
Very positive	Joy

Table 7.1: Applause classification output and emotion correspondence

Emotion communication research strongly focuses on how the emotions are reflected in different parts of our body. When it comes to expressing emotion with facial features, a well defined standard exists called Facial Action Coding System (FACS), developed by Ekman and Friesen [50]. However, no such method can be found for body features. An attempt is made in [40], where authors present a Body Action and Posture Coding System (BAP) based on 141 human body features that can convey emotion.

Alternatively, in [29] Darwin’s observations about human behaviour are generalised and some relevant features are extracted for differentiating behaviours: posture height, shoulder height, arm position, gaze, body activation, head activation

(nodding or shaking), periodicity and exaggeration (range of motion exhibited by each DoF). Bretan et al. [29] applied some of those features in the robot Shimi, in which a dimensional model of emotions is implemented using *Arousal* and *Valence* as reference values to raise different emotional states.

Several research work can be found in the literature regarding humanoid robots showing emotions. Focusing on those related to the emotional behaviour of the robot NAO, we find interesting the work developed by Erden [52]. He proposes a set of body postures based on the human body model described in Coulson’s [38] work to display different emotional states, such as angry, happy and sad. Another example is the approach taken by Barakova and Lourens [16], in which a framework for expressing and interpreting emotional movements based on the Laban Movement Analysis is presented. They propose a method and language for describing, visualising, interpreting and documenting human movement.

We chose to represent our Bertsobot’s emotional behaviour by means of gestures, i.e., fluent sequences of postures. For each emotion class defined in Table 7.1, a set of 3 predefined gestures has been prepared, giving a total amount of 12 different gestures. Each gesture consists of a fluent concatenation of postures. Those gestures have been generated based on the relation between posture and emotional state described in [100] and reproduced here in table 7.2.

Emotion	Frequent posture features
Anger	Head backward, chest not backward, no abdominal twist, arms raised forwards and upwards, shoulders lifted
Joy	Head backward, chest not forward, arms raised above shoulder and straight at the elbow, shoulders lifted
Sadness	Head forward, chest forward, no abdominal twist, arms at the side of the trunk, collapsed posture
Surprise	Head backward, chest backward, abdominal twist, arms raised with straight forearms
Pride	Head backwards or tilted lightly, expanded posture, hands on the hips or raised above the head
Fear	Head backward, no abdominal twist, arms raised forwards, shoulders forwards
Disgust	Shoulders forwards, head downwards
Boredom	Collapsed posture, head backwards not facing the interlocutor

Table 7.2: Posture features of emotions

Therefore, after the classification of a feedback event, the corresponding emotion is selected, and as response one gesture is randomly chosen out of the corresponding set and displayed to the audience. Figure 7.3 shows several examples of those prepared emotional reaction gestures.

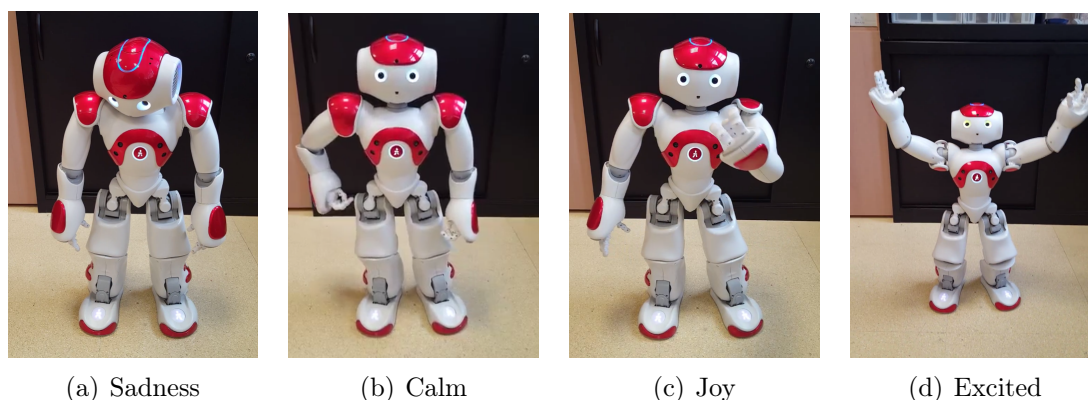


Figure 7.3: Some examples of emotional reaction gestures

### 7.2.3 Generating verses with sentiment

The purpose of the verse is not only to convey a message in the form of a poem but also to respond to an affective target and/or to create an affective response in the audience. The emotion perception system described in Section 7.2.1 is intended to be used as input in the verse generation process. In this way, BertsoBot will be able to use audience’s feedback for self-evaluation and give a response accordingly, maintaining the affective response when the audience reaction is positive, and changing the sentiment target when it is not.

*Bertsolari-s* do not always respond to the proposed theme and context with the same mood. Sometimes their response to a specific theme is positive, while others they address it from a negative perspective. We also want to find a way to introduce emotions in our verse generation system.

The verse generation system (see Section 5.1.2) only took as input the exercise type and the topic or four rhymes (depending on the exercise type). A new input has been added to the system, the sentiment polarity. In order to compose verses with a predetermined sentiment polarity (positive, negative or neutral) a sentiment tool has been developed and integrated in the verse generation system. Such tool is used in the sentence selection process allowing to choose those sentences that match the intended sentiment. To extract the sentiment evaluation from the sentences, we use EliXa, a supervised Sentiment Analysis system [137]. It estimates the negative, neutral and positive sentiment values in short texts by means of a multi-class Support Vector Machine (SVM) algorithm. See Astigarraga’s research work [6] for a deeper explanation about the verse generation system.

Figure 7.4 shows the current version of the verse generation system. In addition to the type of exercise and the topic or rhymes, the system also receives as input the affective state (sentiment polarity) with which the robot must compose the verse.

The verse generation system is now connected with the applause classification system. If the sentiment polarity is not given by the emcee, the system generates the first verse with a neutral mood by default. After that, the applause classification system is used for self-evaluation maintaining the affective response when the



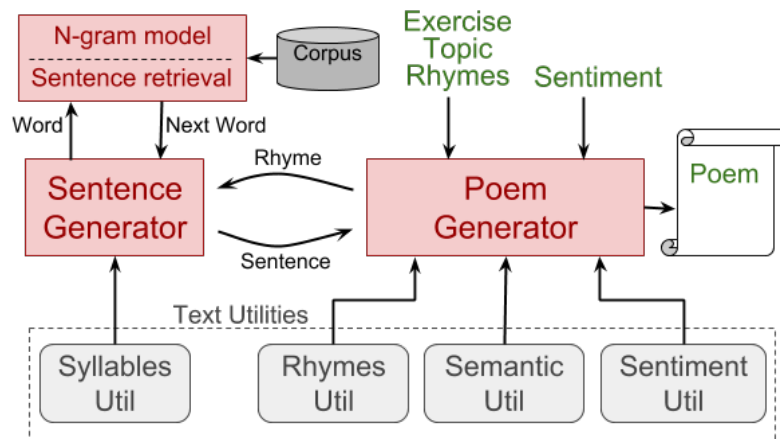


Figure 7.4: Emotional Verse generation system

audience reaction is positive, and changing the sentiment target when it is not.

### 7.3 Talking with sentiment

Speaking in monotone, without any kind of expression is not natural if we intend a natural interaction between human and robots. While speaking, a social robot has to generate credible body language and talk with life and expression, i.e. with emotion, in order to be socially accepted.

Body language is an important mean of communication; face expressions, body gestures, postures, and movements are used to convey information about the emotions and thoughts of the sender while supporting verbal communication. In other words, is the key to express emotions.

Several works propose facial expressions as principal mechanism to show emotions. Kismet [26], Nexi [28], and Sophia [133] are three well-known examples of robotic heads endowed with several facial features (such as eyebrows, eyelids, eyes, mouth, lips, etc.) able to show high facial expression capabilities. There are also other possible facial communicative channels for robots that do not have those features, for instance, colour LEDs. Low-resolution RGB-LEDs can evoke associations to basic emotions (happiness, anger, sadness, and fear), by using suitable colours [86] and combining colours with dynamic light patterns [56].

Relevant work can also be found related to the body expression of NAO and Pepper. Beck et al. [17] investigate the creation of an Affect Space, based on the VAD model, for the generation of emotional body language to be displayed by robots; they assessed the effect of varying a robot's head position on the interpretation of predefined emotional key poses. Tielman et al. [158] define a model for a expressive behaviour of the NAO robot in which *Valence* and *Arousal* values are influenced by the emotional state of its interaction partner and emotional occurrences while interacting with its environment. NAO expresses emotions through its voice, eye

colour, and predefined posture and gestures. Deshmukh et al. [43] modulates some default animations of Pepper by systematically varying the amplitude and speed of the joint motions and gathering user evaluations of the resulting gestures. Finally, Claret et al [36] presented a model that for a given emotion, generates specific kinematic motions of the Pepper robot. The emotion is defined as a point in VAD space, and the desired emotion values are transformed into the kinematic features jerkiness, activity and gaze, which are mapped to a continuous range of body configurations.

Naturally, speech also plays a relevant role to convey emotions, and human voice can be shaped in a very complex way. In the context of human-robot interaction, Crumpton and Bethel explain the importance of using vocal prosody in robots to convey emotions [39].

Our approach combines all the previous aspects in a expressive behaviour that endows the robots with the ability to adapt their way to express different emotions, defined in the VAD space, according to the sentiment of the speech. Head and arms movements, along with eye LED lighting and voice intonation are combined to generate an adaptive expression, i.e., to make the robot express the content of a spoken text with emotion. For the time being we only consider to express the emotion in the sadness-happiness continuum, considering the midpoint as neutral. In order to make the robots able to express the emotional content of a spoken text, two main tasks are required: extract the emotion from the text and translate it to a robot expression. This process can be summarised into three main steps:

1. Extract the sentiment from the text. A sentiment analyser assesses the sentiment of the text and gives as output a descriptor with information about the polarity of the sentiment (positive/negative/neutral).
2. Sentiment to emotion conversion. In this step, the sentiment polarity is encoded into emotion. Only sadness, neutral and happiness emotions have been considered in this work.
3. Generate the appropriate expression. The translation from emotion to expression is performed in this stage. The robot shows an emotional expression by means of body expression (talking gestures), facial expression (eyes lighting) and voice intonation (pitch and speed variation).

In the following sections the approach taken to perform 1-3 steps will be described in detail.

### 7.3.1 Text sentiment extraction

Sentiment analysis is the research field related to the analysis of people's opinions, sentiments, evaluations, attitudes, and emotions from written language [121]. The main purpose of sentiment analysis is to extract the polarity (positive/negative/neutral) of a given text, but more advanced sentiment analysers appraise the emotional content from the text as emotional state in the VAD space.

Bertsobot already includes a sentiment analyser in the verse generation system (see Section 7.2.3). EliXa is used to extract the sentiment polarity from the sentences

in order to generate verses (in Basque) with an affective state. Now, we intend to generalise the sentiment analysis and integrate it into the global talking behaviour, a fundamental capability for social robots. This step forward is given by developing a “Sentiment Analyzer” module that: on the one hand, extracts the sentiment polarity of a given text, and on the other hand, appraises the emotional content from the text as emotional state in the VAD space.

To that end, we propose a sentiment analyser module that combines two different open-source tools: EliXa [137] and MixedEmotions<sup>1</sup> sentiment analysers. EliXa extracts the sentiment polarity from the text but without any confidence level, and it is available in Basque, Spanish and English. MixedEmotions assesses a confidence level for each of six basic emotions (surprise, anger, disgust, fear, sadness and happiness), and it is available in Spanish and English. The module developed combining these two tools outputs a descriptor composed by the sentiment polarity label extracted from the text using EliXa and the VAD numbers obtained with MixedEmotions. As MixedEmotions tools are not available for Basque, the translation of the text from Basque to English has to be performed in advance.

### 7.3.2 Sentiment to emotion conversion

From the two pieces of information returned by the text sentiment analyser module we want to assess the emotion of a text in the sadness-happiness continuum, computing its numerical value.

As mentioned before in this chapter, in dimensional affective models emotions are described as linear combinations of *Valence-Arousal-Dominance*, and *Valence* defines how positive or negative the stimulus is. Therefore, the *Valence* deals with the positive or negative character of the emotion, which scales from sadness to happiness.

The “Emotion Selector” module we have developed to compute the conversion from sentiment to emotion uses the output obtained from the “Sentiment Analyser” module. We have defined the [0,10] range for the numerical value that represents the emotion, and it is divided into three parts; where the interval [0,4.5) corresponds to negative polarity, [4.5,6.5] to neutral polarity and (6.5,10] to positive polarity. This division of the scale comes from the observation of the *Valence* values returned by MixedEmotions in a set of sentences. The aim is to directly translate the *Valence* into the sadness-neutral-happiness scale with a few caveats. The approach employed is the following one: first we analyse the polarity of the text according to EliXa, and if the *Valence* according to MixedEmotions lies in the interval corresponding to the polarity (where lowest interval corresponds to sadness emotion, middle interval to neutral emotion and highest interval to happiness emotion), then the *Valence* value is used as it is. Otherwise, the limit value of the closest interval is chosen.

For the time being, the emotion appraisal is done as a direct translation from the sentiment value into emotion in sadness-happiness continuum. It is worth mentioning that more inputs and more emotions should be considered for the emotion

---

<sup>1</sup><http://mixedemotions.insight-centre.org/>

appraisal in the future.

### 7.3.3 Displaying the emotion

Once the analysis of the emotion is done, the robot must appropriately express such emotion. Our approach to display emotions consists of mapping the emotion into expression that combines natural body gestures enriched with facial expressions and voice intonation. The duration of the expression will be fixed by the duration of the speech (generated audio file duration).

#### Body gestures

Typically, humans accompany their speech with body gestures (head, hands and arms movements). We use gestures to communicate with others; gestures are used both to reinforce the meaning of the words and to express feelings through non-verbal signs.

Consciously or not, emotions are reflected in different parts of our body [100]. For instance, the position of the head convey sadness if is tilted down, or happiness if is tilted up. On the other hand, we do not stay still while talking, we naturally gesticulate with hands. McNeill defines gestures as the movement of the arms and hands which is synchronised with the flow of the speech [113]. He categorises five main types of conversational gestures (illustrators): emblems, deictic, iconic, metaphoric and beats; emblems are those gestures whose meaning can be understood without spoken words; deictic gestures utilise arm and hand movements to direct the listener's to specific event or object in the environment; iconics refer to gestures representing concrete things and actions; metaphorics describe the content of abstract ideas; and beat gestures are rhythmic hand motions that move up and down in synchrony with the speech. Unlike the others types, beats are not associated with particular meanings, and they occur with the "rhythm" of the speech. Such kind of gestures have been considered both in the work related to this chapter and with that of Chapter 8.

The talking body language described in Section 5.3.4 did not show any emotional state; the sequence of the predefined gestures to be executed was randomly selected and the number of gestures was chosen according to the duration of the speech. Here on, the strategy to generate talking gestures is similar, but it modifies the gesture selected in two ways: changing the head tilt angle and the execution velocity of the arms movements. For the head tilt, the *Valence* obtained is directly translated from VAD space to head pitch physical range. Instead, for the execution velocity of the arm movements, we have set a maximum and minimum range and just like with the head, the *Valence* obtained is translated to minimum-maximum velocity range. In this way, if the emotion to be shown is "happiness" the gesture will be executed at a faster pace than when the gestures bound to the emotion "sadness". These changes have been introduced in the "Body Expression" behaviour previously developed.

### Facial expression

The design of humanoid robots' eyes is usually inspired by human face, trying to exactly reproduce human eyes' shape and movements. However, SoftBank's robots have some limitations due to the structure of their eyes. In particular, NAO and Pepper robots' eyes are composed by two rings of LEDs with a black pupil inside. The LEDs can be controlled to show different hues, to change colour intensity and can be turned on/off for different time duration.

Johnson et al. [86] demonstrate in their work that NAO's eyes can be used to express emotions. Taking inspiration from their colour-emotion study, in our approach we adopt the same colour configuration, and in addition we use the emotion *Valence* to change the intensity of the colour. For each emotion a range of colour in the RGB space has been defined. Sadness is displayed by a dark blue-greenish colour that varies from RGB(0, 0, 255) to RGB(0, 255, 255), neutral is displayed by a light blue-white colour from RGB(127, 255, 255) to RGB(255, 255, 255) and happiness is displayed by a yellow colour from RGB(76, 76, 0) to RGB(255, 255, 0).

The "Eyes Lighting Controller" is employed to convert emotion into facial expression, specifying the colour and the intensity of each eye LEDs (see Figure 7.5). The controller codes the emotion's *Valence* value code into the RGB space that will be displayed in the robot.

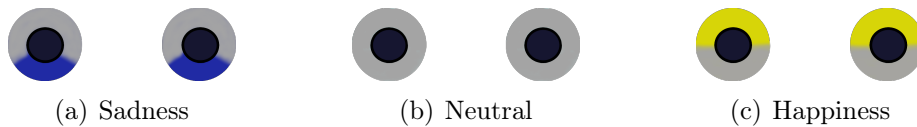


Figure 7.5: Sadness-Neutral-Happiness emotions displayed by the robot's eyes

### Voice intonation

In ordinary life humans use different voice intonation depending on the context and also to emphasise the message being conveyed. The voice intonation has a key role to understand the mood of the speaker. The influence of the voice intonation in emotional expression is clearly argued in [15]. The authors prove that some emotions, such as fear, happiness and anger, are portrayed in a higher speech rate and also at a higher pitch than emotions such as sadness.

We have used the happiness, neutral and sadness intonations to portray the three emotions available in our system. For Basque language we have used the AhoTTS synthesizer, that offers different types of voice intonation; one for each of six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) and another one for neutral emotion. AhoTTS is also available for Spanish and English, but unfortunately the voice intonation option is not possible for those languages. Therefore, for Spanish and English NAO and Pepper's TTS tool has been used, which employs the ACAPELA<sup>2</sup> speech synthesizer. This synthesizer does not offer

<sup>2</sup><http://www.acapela-group.com/>

direct voice intonation selection as AhoTTS does, but it does provide the option to setup some voice parameters, such as pitch and speech rate, which can be tuned to obtain different voice intonations.

Our approach consists of changing the pitch and speed rate parameter values according to the emotion *Valence* value, i.e. the emotion's *Valence* obtained from the emotion appraisal is normalised between the maximum and minimum values for the voice pitch and speed rate. Maximum and minimum values have been experimentally defined for our system. This approach is employed for Spanish and English. For Basque, the intonation parameters are fixed and only the emotion type can be selected though.

## 7.4 Contributions and conclusions

Two emotional modules have been developed that contribute to improve the overall robot behaviour.

On the one hand, the first module endows the robots with the ability to understand and reacting to the audience feedback. Such behaviour permits the robot to infer self-evaluation about its actions and to close the loop of the performance; perceiving audience's feedback by measuring the intensity and the duration of the applause, and adapting its response to better please the public by reacting through subtle gestures and changing the sentiment of the verse.

That behaviour has been used and tested in two public performances. In February 2016 we organised a local event in the Faculty in order to be able to evaluate the applause classification and the emotional expression modules. NAO robot in front of the audience sang previously generated verses staging only the phase of thinking and singing the verse. After each sung verse the robot reacted giving an emotional response according to the audience's applause feedback (see video<sup>3</sup>). In September of the same year Bertsobot was invited to the closing of a summer university course entitled "educational assessment: unresolved matter". In such event, NAO showed its dialogue, body expression and singing capabilities, and it also showed the ability to perceive audience's feedback and to adapt its emotional response accordingly.

The following publication collects the work done related to the behaviour mentioned above:

- **Minstrel robots: Body language expression through applause evaluation.** F. Kraemer, I. Rodriguez, O. Parra, T. Ruiz, E. Lazkano 2016 IEEE-RAS International Conference on Humanoid Robots (Humanoids), Cancun, Mexico, November 2017, pp. 332-337.

On the other hand, the emotional behaviour of the second module provides the robot with the necessary skills to adapt its way to express different emotions while talking according to the sentiment of the speech. The adaptive emotional system we have developed generates an expression combining head and arm gestures along

---

<sup>3</sup><https://www.youtube.com/watch?v=SdxNgmV3CzA>

with eye LED lighting and voice intonation. This “Adaptive Talking Behaviour” is composed by several ROS modules, as illustrated in Figure 7.6. The “Sentiment Analyser” analyses the sentiment of the text, the “Emotion Selector” converts the sentiment into an emotion, the “Eyes Lighting Controller” manages the eyes colour, the “Speech Synthesizer” tunes voice parameters, and the “Gesture Adapter” generates the body expression including arms, hands and head.

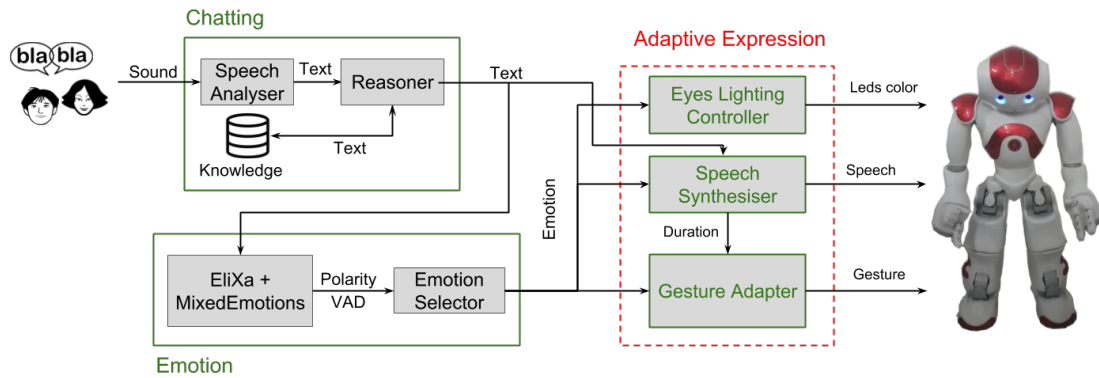


Figure 7.6: Description of the Adaptive Talking Behaviour architecture

This behaviour has not been tested in any public performance yet, but we have recorded a video that demonstrates how our NAO is able to adapt its way to express different emotions while talking according to the sentiment of the speech (see video<sup>4</sup>).

The work related to the “Adaptive Talking Behaviour” is collected in the following publication:

- **Adaptive Emotional Chatting Behavior to Increase the Sociability of Robots.** I. Rodriguez, J.M. Martínez-Otzeta, E. Lazkano, T. Ruiz. 2017 International Conference on Social Robotics (ICSR), Tsukuba, Japan, November 2017, pp. 666-675.

A global overview of all the behaviours developed up to this point is sketched in Figure 7.7. It shows the current state of the system and describes the inter-communication between all behaviours presented up to now. New modules added are highlighted in the figure with red colour, while adapted modules are highlighted with orange colour.

The two emotional behaviours proposed in this chapter have allowed us to improve the social capabilities of our robots. In this way the expression ability of the robots has been considerably enhanced achieving in turn a more natural interaction with the user. Both behaviours have been successfully integrated in Bertsobot’s architecture, however, there are still some limitations we need to consider:

- Talking gestures generation: the strategy employed to generate talking gestures (adapting head tilt and the execution velocity of the arms) has brought

<sup>4</sup>[https://www.youtube.com/watch?v=wI2BD4j4\\_tU](https://www.youtube.com/watch?v=wI2BD4j4_tU)

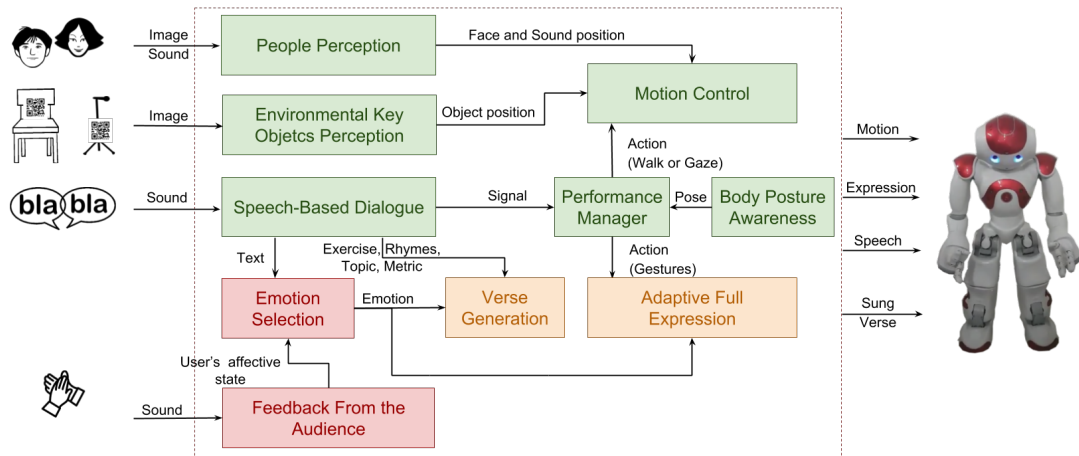


Figure 7.7: Bertsobot's architecture after including emotions into the system

improvements concerning the expressiveness of the robots, but the shortage amount of predefined gestures included in NAOqi's animation library still makes the robot expression ability limited and repetitive.

- Sentiment analyser and emotion selector modules: in many cases there were inconsistencies between the outputs obtained with EliXa and MixedEmotions; high *Valence* values of the emotional state returned by MixedEmotions should match the positive polarity returned by EliXa, or low *Valence* values with negative polarity, but it was not always the case. This made it difficult the conversion from sentiment to emotion. We tried to tackle the problem giving more weight to EliXa's assessment (as explained in Section 7.3.2), but the performance of this module does not fully meet our requirements.



# Chapter 8

## Enhancing spontaneity

This chapter aims to improve the verbal and non-verbal aspects of the robots' talking behaviour that showed insufficient:

1. Talking illustrators (beats): the naive talking gesticulation behaviour was appropriate for the Bertsobot because the talking periods did not take long, the troubadours only may require brief clarifications from the emcee. But we moved forward beyond the Bertsobot and developed a more general talking behaviour that included some emotional aspects that intends to be used for basic social interactions. Thus, the problem of generating repetitive movements becomes more noticeable in longer conversations.

The taken approach relies on the use of generative learning methods and more specifically, the use of Generative Adversarial Networks that produce synthetic gestures (head, arm and hand movements) using proprioceptive 3D joint information exclusively.

2. The second improvement goes hand in hand with the modification of the sentiment extraction system. The EliXa and MixedEmotios tools combination previously presented pop up contradictory results that were not solved and encouraged us to look for new alternatives. In this vein, we present a new version of the text sentiment extractor that uses VADER [79] as a unique tool.

### 8.1 Generative models and their applications

Generative models are probabilistic models capable of generating all the values for a phenomenon. Unlike discriminative models, they are able to generate not only the target variables but also the observable ones [153]. They are used in machine learning to (implicitly or explicitly) learn the distribution of the data for generating new samples. There are many types of generative models. For instance, Bayesian Networks (BNs), Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) are well known probability density estimators. Explained in short, a BN is

a model representation for reasoning under uncertainty [51]. Formally, its representation is a directed acyclic graph where each node represents a random variable and the edges represent dependence relations between them; GMMs are those that attempt to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset [53]; and HMMs can be considered a generalisation of mixture models where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other [129].

Deep learning techniques have also been applied to generative models, giving rise to deep generative models. A taxonomy of such models can be found in [67]. Here we will focus on Generative Adversarial Networks (GANs) [68], deep generative models capable to implicitly acquire the probability density function in the training data, being able to automatically discover the internal structure of datasets by learning multiple levels of abstraction [98].

Applications of generative models range from photo-realistic single image super-resolution [99] and text-to-image synthesis [132] to handwriting sequences generation [69] using recurrent neural networks (RNN) or speech synthesis [162] based on WaveNet [161], an autoregressive deep generative model. In [127] the authors propose Deep Generative Spatial Models (DGSM), the first application of Sum-Product Networks to the domain of robotics. A generative model is able to learn a single, universal model of the robot’s spatial environment. In astronomy GANs are becoming popular for improving images [139].

Generative models are also being used for motion generation. In [94] the authors propose the combination of Principal Component Analysis (PCA) [171] and HMMs for encoding different movement primitives to generate humanoid motion. Tanwani [153] uses HSMM (Hidden Semi-Markov Models) for learning robot manipulation skills from humans. Focusing on social robotics, some generative approaches are being applied with different objectives. In [108] Manfrè et al. use HMMs for dance creation and in a later work they try variational auto-encoders again for the same purpose [11]. Regarding the use of adversarial networks, Gupta et al. [72] extend the use of GANs to generate socially acceptable motion trajectories in crowded scenes in the scope of self-driving cars.

### 8.1.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [68] are semi-supervised emerging models that basically learn how to generate synthetic data from the given training data. A GAN network is composed by two different interconnected networks. The *Generator* ( $G$ ) network generates possible candidates so that they are as similar as possible to the training set. The second network, known as *Discriminator* ( $D$ ), judges the output of the first network to discriminate whether its input data are “real”, namely equal to the input data set, or if they are “fake”, that is, generated to trick with false data. As one of the most researched field of application of GANs is in the context of image generation, the GAN generator is typically a deconvolutional neural network, while the discriminator is a convolutional one. The general architecture of this type

of GAN with the  $G$  and the  $D$  networks is shown in Figure 8.1. However, when the input data has not a spatial structure that makes convenient the use of convolutional layers, standard dense layers are used instead. The latter approach is the one used in this research.

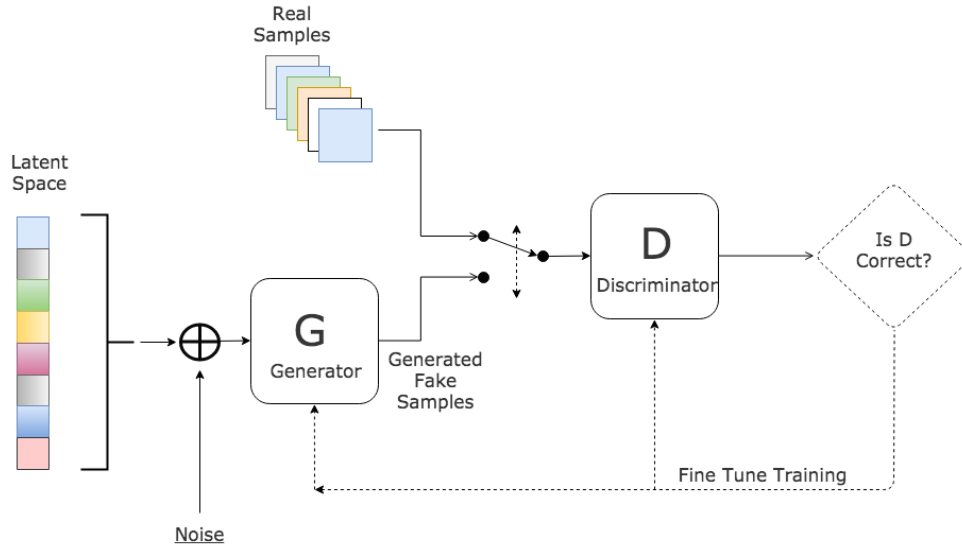


Figure 8.1: Description of a general GAN architecture

In the first step,  $D$  takes as input both, real data and fake data, and returns for each sample its probability to come from real data. In the second step, the  $G$  network is trained. While the parameters of  $D$  are fixed, in each epoch, the weights of the  $G$  network are updated to let the discriminator results on the sample generated by  $G$  be as near as possible to 1. That is, this second step is aimed to modify the  $G$  network in order to be able to generate samples that can trick the  $D$  network.

The  $G$  network is never exposed to real data, therefore the only manner to enhance its generation capability is through the interaction with  $D$  by means of the output. Instead,  $D$  has access to both, real data and fake data, and produces as output the ground truth to know if the data came from the generator or the dataset. The discriminator's output value is exploited by the generator to enhance the quality of the forgery data.

Back-propagation [74] is applied in both networks to enhance the accuracy of the generator to produce valid movements; on the other side, the discriminator becomes more skilled to flag false data.

## 8.2 Beyond deterministic body language

According to Beck et al. [18], there are three main motion generation approaches: manually creating motion, motion capturing, and motion planning; for manual creation, it is required to set each joint position of the humanoid robot for each key

frame (time step); the motion capture-based approach tries to mimic human gestures, recording human movements and mapping these data to a humanoid robot; and motion planning approach relies on kinematics and/or dynamics equations to solve a geometric task. They find that motion capturing approach produces the most realistic results, because the robot reproduces previously captured human movements. They also point out that motion planning approach is the only one that generate gestures that can be adapted to new situations, but the challenge here is how to use these methods to generate believable movements for social interaction.

With respect to the talking gesticulation behaviour we are focused on, in its current state (see Section 7.3), we randomly selected the gestures (full animations) from a set of movements previously compiled, generating sequences of gestures to accompany the speech. Those movements were manually generated, by means of the *Choregraphe*'s animation creation tool. Each joint position was recorded at a precise key frame, generating realistic movements that resemble to those gestures shown by humans.

The number of gestures selected was determined by the duration of the speech. However, despite the attempts made to improve the body language capabilities of our robots, and after observing the behaviour of the robot when talking, we concluded that such approaches were prone to produce repetitive movements, resulting in unnatural jerky behaviour.

What if we could generate novel movements each time the robot talks? The automatic generation of non predictable motions would undoubtedly enhance robot's spontaneity. Nevertheless, the motion producing system should guarantee the absence of weird ticks and too fast moves that would be categorised as unnatural and strange by any human observer.

We started to investigate whether generative models and GANs in particular could be used to create novel movements while retaining the nature of the movements that we already have. The goal is to provide the GAN network with 3D proprioceptive sensor information to generate natural talking gestures. The gesture repertoire of the robot will be limited to head motions and beat gestures [113], i.e., rhythmic arm and hand motions, that move in synchrony with the speech.

With that aim, a GAN was trained with the information obtained from Pepper's proprioceptive sensors, specifically it takes as input uniquely motors joint position information. In order to collect training data, we sampled the *poses* during the selected animations with a frequency of 4 Hz. Pose refers to the fixed picture of the position and orientation of the joints of the robot. A pose is composed by 14 float numbers and thus represented by a set of 14 joint values, comprising robot's head, hands and arms (Table 8.1). Pelvis, knee and wheeled base information were ignored because those elements are not involved in talking adaptors. As we are interested in generating movements, i.e., a sequence of poses, the input to the learning process needs to take into account the temporal sequence of poses. We defined the *unit of movement* as a sequence of four consecutive poses, i.e. a vector of 56 float numbers, 14 from each pose and four poses concatenated (see Table 8.2). The training set is thus a set of units of movement generated taking four consecutive poses from a database of poses. The output of the generative model will be a unit a movement

as well.

Head	$H_\gamma(J_1)$	$H_\beta(J_2)$
Right Shoulder	$RS_\beta(J_3)$	$RS_\alpha(J_4)$
Right Elbow	$RE_\beta(J_5)$	$RE_\gamma(J_6)$
Right Wrist	$RW_\gamma(J_7)$	
Right Hand	$RH(J_8)$	
Left Shoulder	$LS_\beta(J_9)$	$LS_\alpha(J_{10})$
Left Elbow	$LE_\beta(J_{11})$	$LE_\gamma(J_{12})$
Left Wrist	$LW_\gamma(J_{13})$	
Left Hand	$LH(J_{14})$	

Table 8.1: Format of a robot pose.  $\alpha$ : roll,  $\beta$ : pitch,  $\gamma$ : yaw angles. Hands can be opened or closed. In parenthesis, a more convenient notation for formal use, where  $J_i$  refers to joint of index  $i$

$$J_1(t) \cdots J_{14}(t), J_1(t + \Delta t) \cdots J_{14}(t + \Delta t), \dots, J_1(t + 3\Delta t) \cdots J_{14}(t + 3\Delta t)$$

Table 8.2: Characterisation of a unit of movement (4 consecutive poses).  $\Delta t$  depends on the data sampling frequency

The talking gestures generation system executes those units of movement produced by the generative model. It needs to be mentioned that the temporal length of the audio intended to be pronounced by the robot determines the number of units of movement required to the generative model. Thus, the execution of those units of movements, one after the other, defines the whole movement shown by the robot.

These videos demonstrate the appropriateness of the obtained behaviour:

1. A first video<sup>1</sup> shows the evolution of the robot behaviour during different steps of the training process. The final number of epochs has been empirically defined, observing the behaviour of the robot.
2. A second video<sup>2</sup> qualitatively demonstrates the adequateness of the approach by showing how the robot behaves while talking.

Observed robot behaviour suggests that GANs are a suitable method for generating robot movements that capture the essence of the predefined gestures, while allowing more variability and, overall, giving a subjective impression of naturalness.

### 8.2.1 Comparing GAN with other approaches

GANs are deep learning methods and as such, they need high computational resources to be trained. A question that evidently arises is why we chose such a

<sup>1</sup><https://www.youtube.com/watch?v=AW3BmfS7DIY>

<sup>2</sup><https://www.youtube.com/watch?v=KVyTbFEMcHE>

method and if it wouldn't be enough to employ a less demanding generative approach. To try to answer that issue, we compared the obtained motion generation system with three other generative approaches:

1. Random generation of movements, by just randomly concatenating poses from already existing ones.
2. Gaussian mixture models.
3. Hidden Markov models.

Evidently, it is not easy how to make that comparison. We are dealing with rather subjective properties of robots' behaviour like naturalness and spontaneity difficult to be quantitatively measured.

We made a two step performance analysis. On the one hand, variability decomposition after applying Principal Coordinate Analysis and the projections of the joints on the corresponding principal axes for each method revealed that GAN retains the underlying structures of the joints present in the training database. Other methods do not show this property. On the other hand, we defined a method to analyse the robot motion based on three factors: Norm of jerk, 3D space coverage and length of the generated paths. Gestures generated using GAN showed smaller jerk values, together with smaller path lengths but at the same time, the corresponding dispersion measure value is the highest for both, left and right hands. Obtained results show that GAN produces trajectories more similar in shape to those in the original gesture set.

A third video<sup>3</sup> shows the differences among the distinct methods in the robot.

### 8.3 Combining the gesture generation system with the Adaptive Talking Behaviour

It is important to adapt the gestures to convey emotion, introducing some variations according to the robot's "mood". Recalling Chapter 7, to change the motion timing of the gestures and the head tilt, the sentiment of the text to be pronounced was extracted by means of EliXa and MixedEmotions. However, several issues aroused from that combination. They often gave contradictory results that were not easy to balance. In this step, we propose a new version of the emotional talking behaviour by changing the sentiment analyser. We chose to use VADER [79] instead, because it provides both, the sentiment polarity and the compound score, and it is no more necessary to combine two different tools.

This VADER sentiment analyser is based on dimensional affective models, and gives an output composed by:

1. The score ratios for proportions of text that fall in each category.

---

<sup>3</sup>[https://www.youtube.com/watch?v=YEG326L\\_p4s](https://www.youtube.com/watch?v=YEG326L_p4s)

2. A compound score, obtained by summing the *Valence* scores of each word in the lexicon.

The VADER python library<sup>4</sup> is mainly available for English, but it also can work with texts in other languages; so for, it provides an option to automatically translate the text into English by using a web service call.

Heading back to the “Adaptive Talking Behaviour”, the new “Sentiment Analyser” module takes as input a text and a language and first, makes a translation to English (if required), then infers the sentiment polarity (positive/negative/neutral) with the compound score obtained from VADER, and finally gives as output a descriptor with the sentiment polarity label and the compound score obtained. This time, the approach employed in the “Emotion Selector” lies on a direct translation of the compound score into the sadness-happiness continuum in the *Valence* axis that ranges from  $[-1,+1]$  (Sadness: compound score  $\leq -0.5$ ; Neutral:  $-0.5 < \text{compound score} < 0.5$ ; Happiness: compound score  $\geq 0.5$ ).

The talking gestures generation system developed using GANs has been successfully integrated in the architecture. Main changes introduced refer to the “Adaptive Talking Behaviour”, in which the previous “Gesture Adapter” module has been replaced by the “Gesture Generator” module (the system presented in this chapter). Figure 8.2 shows the architecture of the “Adaptive Talking Behaviour” after integrating the GAN-based talking gestures generation system and the VADER based emotional system. Changes are highlighted in orange colour.

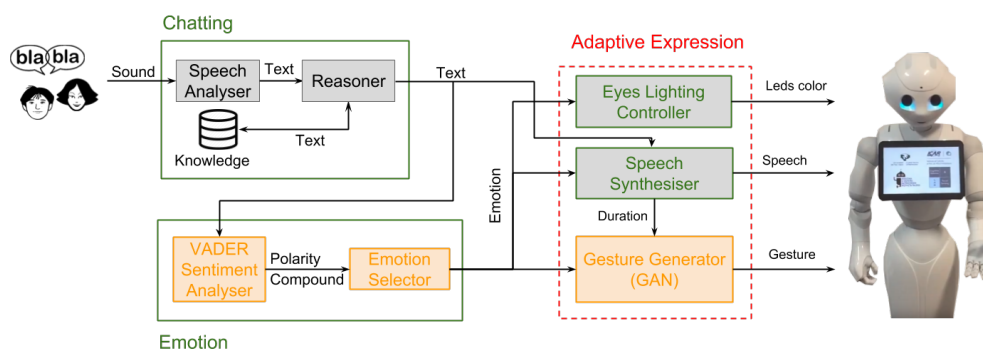


Figure 8.2: Description of the Adaptive Talking Behaviour architecture after integrating the GAN-based gesture generation and the new sentiment analyser

The potentiality of the generated robotic gesticulation is enhanced by the possibility to convey also emotions depending on the pronounced speech, and enriching the interaction with other relevant factors such as head and arms movement speed, the tone of the voice and the colour of the eyes. The robot expression is now displayed through generated beats adapted and combined with eyes’ colours and voice intonation.

<sup>4</sup><https://github.com/cjhutto/vaderSentiment>

## 8.4 Contributions and conclusions

Two are the contributions of the work described in this chapter:

1. The gesture generation system developed for talking illustrators:
  - **Spontaneous Talking Gestures Using Generative Adversarial Networks.** I. Rodriguez, J.M. Martínez-Otzeta, I. Irigoien, E. Lazkano. IEEE Robotics and Automation Society, 2018 (Submitted).
2. The enhancement of the text sentiment extraction procedure and its application to modulate the gestures generated for talking:
  - **Talking with Sentiment: Adaptive Expression Generation Behavior for Social Robots.** I. Rodriguez, A. Manfré, F. Vella, I. Infantino, E. Lazkano. 2018 International Workshop of Physical Agents (WAF), Madrid, Spain, November 2018 (Submitted).

We developed a suitable generator of rhythmic gesticulation movements with few other expressive features (the head position, the arm movements velocity, the variation of the voice tone and the change of colour eyes) dependent from sentiment detected on sentences. We have not had the opportunity to test this behaviour in any public performance yet, but we have recorded a video<sup>5</sup> that demonstrates how our Pepper is able to adapt its way to express different emotions while talking according to the sentiment of the speech.

Robot behaviour observed suggests that GANs are a suitable method for generating robot movements that capture the essence of the predefined gestures, while allowing more variability and, overall, giving a subjective impression of naturalness. The appropriateness of GANs to cope with this type of problem in which robot motion is generated using proprioceptive 3D joint information has been demonstrated by making a comparison with other (non-deep) generative approaches.

---

<sup>5</sup><https://www.youtube.com/watch?v=Vo0t9l4exwI>



## Chapter 9

# A framework for socially interacting robots

We adopted *bertsolaritza* as ecological niche and used it as starting point to develop the building blocks of a control framework for social robots. We focused on creating a practical control architecture that follows the dynamics of real events, as verse-makers do:

1. Wait sitting for its turn.
2. When demanded, place itself in front of the microphone and listen to the exercise proposed by the emcee.
3. Compose and sing the verse to the public.
4. Observe and receive audience's feedback and react accordingly.
5. Go back to its sitting place.

During several years, we endowed the Bertrobot with different modules that define its behaviour. In its final stage, it supplies the robots with some of the *bertsolari-s'* capabilities allowing them to take part in public performances. Besides singing and chatting, our robots are able to perceive the feedback and emotional state of the audience through their applause and react accordingly, as human oral improvisers do, modifying in real time the sentiment of the poem and its corporal expression accordingly.

All these tasks are accomplished and managed by a ROS based control architecture. The control architecture is composed by different behaviours or modules that make the robot act in a consistent manner. Those ROS modules are activated by different stimuli (speech orders, object detection, etc.) and depending on the state of the performance the robot executes the corresponding task.

This chapter is intended to give a summary of the evolution of the architecture and describe in more detail its final state.

## 9.1 Stages of the architecture

The evolution of the architecture has evolved over time, however we can highlight three main prototypes or versions of BertsoBot's architecture:

1. *Version 1*: the first version of the architecture was implemented in Tartalo and Galtxagorri (two of our wheeled mobile robots) using SORGIN [9], a software framework for behaviour control implementation (coupled with Player<sup>1</sup>). These robots show verse generation (using sentence-retrieval method) and singing capabilities, but they could only perform one exercise, rhymes given. Besides singing, they also incorporated the verbal communication and basic body expression capabilities; robots could listen and talk to the emcee but the dialogue was a precompiled script, and they only made basic movements such as turn on the base and move the camera. Robot movements on stage were limited to teleoperated commands. Figure 9.1 shows the first control architecture.

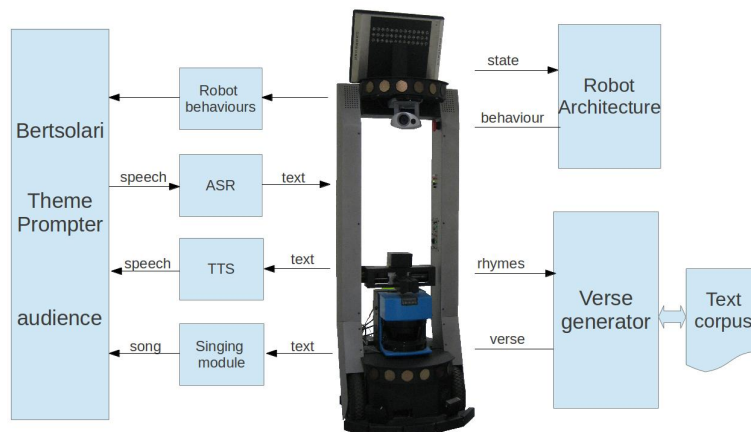


Figure 9.1: Version 1: BertsoBot's architecture implemented in Tartalo and Galtxagorri

2. *Version 2*: the second version of the architecture added new features in respect of the first version. Some significant changes were made: the architecture was reimplemented in ROS and the wheeled mobile platforms were replaced by the humanoid robot NAO. Verbal communication capabilities were improved; a new exercise was included in the verse generation module, theme given, and the dialogue manager was added allowing the robot to maintain a simple conversation with the emcee. In addition a key object recognition (chair and microphone) module was integrated together with autonomous navigation capability. The body expression capability of the robot was also improved, a gesture library to represent different states (waiting, talking, thinking and

<sup>1</sup><http://playerstage.sourceforge.net>

singing) of the performance was included; all those gestures were generated with *Choregraphe*. Figure 9.2 describes the second version of the architecture;

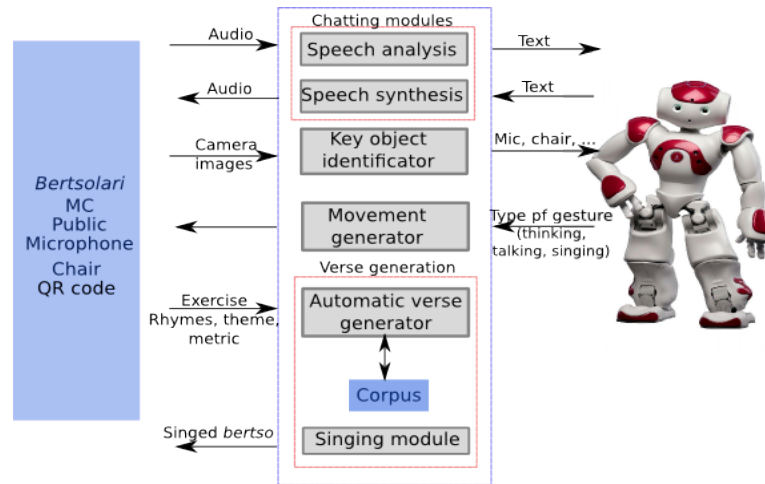


Figure 9.2: Version 2: Bertsobot's architecture implemented in NAO

3. *Version 3*: the latest version of the architecture (working on NAO and Pepper) includes several important changes in relation to the previous versions. Verbal communication capability has been enhanced, on the one hand by the people perception module that enriches the interaction with the user, and on the other hand, by the new features added in the verse generation module, that now includes the option to compose the verse using n-gram generation methods and integrates the affective state in the creation process. Emotions have also been integrated in the architecture: two emotional behaviours have been developed that endow the robots with the capabilities for, on the one hand, understanding and reacting to the audience feedback, and on the other hand to adapt its way to express different emotions while talking. The autonomy of the robot has also been improved by the body self-awareness behaviour, that includes the body posture awareness and the action selection modules. Lastly, in order to enhance robot spontaneity the gesture generation system using GANs has been integrated in the talking behaviour, which now automatically generate gestures based on the sentiment of the speech.

Focusing attention in Figure 9.3 that graphically describes the *Version 3* of the architecture, we would like to emphasise that it is beyond the Bertsobot system. Only the “Verse Generation” and “Feedback From the Audience” boxes betrays its task specificity. We think that robot development is task specific and that social robots must be provided with some general capabilities but their behaviour will be dependent on the niche they are meant to occupy.

In that vein, the generative approach for talking gesticulation and the emotion selection and representation perspective taken are applicable beyond the Bertsobot.

We will focus on describing the different behaviours that form the latter one. The “Performance Manager” is the behaviour that brings the coherence necessary to

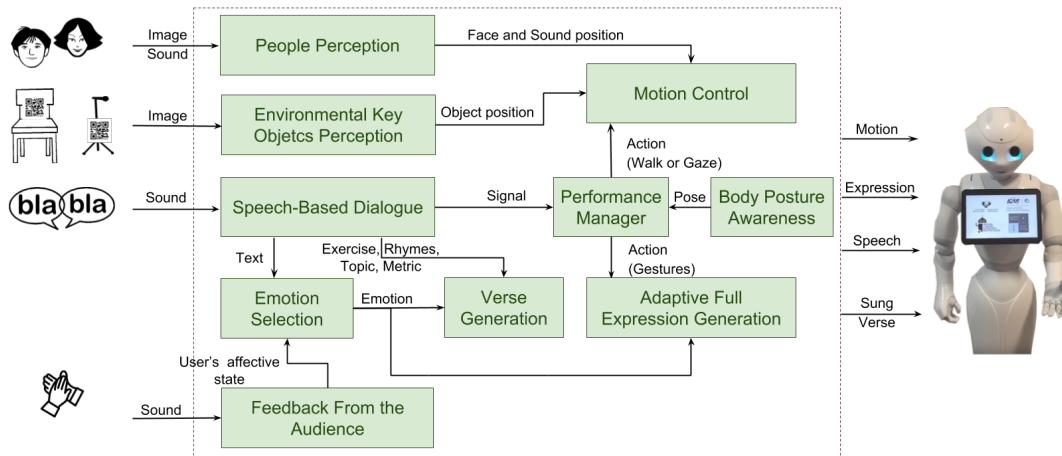


Figure 9.3: Version 3: BertsoBot's final architecture implemented in Pepper

the system in order to follow the dynamic of a performance. It decides the action/actions the robot has to perform depending on the current state of the performance and the current body posture of the robot. The latter information is provided by the “Body Posture Awareness” module. The “People Perception” as well as the “Speech-Based Dialogue” behaviours allow the interaction with the emcee, while “Environmental Key Objects Perception” provides the robot with necessary skills to interact with environmental key objects. These interactions, usually executed as motion actions, are managed by the “Motion Control” behaviour. The verse is composed and sung by the “Verse Generation” process, and audience applauses, which affect the robot’s emotional state, are captured and classified by “Feedback From the Audience” behaviour. The robot’s emotional state is managed by the “Emotion Selection” module, which decides the emotional state that the robot must show considering both the feedback obtained from the audience and the emotion extracted from the text to be said by the robot. The emotional response of the robot is reflected in the verse generation process and in the robot body expression. The robot body expression is managed by the “Adaptive Full Expression Generation” behaviour that decides the expression the robot has to show according to the state of the performance, the robot posture, and the emotional state of the robot. Such expression can be both a predefined gesture chosen from the appropriate gesture set or a full expression generated.

## 9.2 Evolution of the system through public performances

The robots’ performance capabilities have been demonstrated in different events in a 5 years period. These public performances show the evolution of the BertsoBot project since its start up, when no humanoid platform was available, and up to now. First experiments were carried out by Tartalo and Galtxagorri, two of our wheeled

mobile robots. Later on, we move to humanoid robots, and we employed NAO and Pepper robots.

The objective of the live performances was to bring social robotics to the general public and, along with that, to receive audience's feedback about human-robot interaction.

Let's make a quick review of the public performances carried out with our robots:

- **2012/04: First public appearance:** Inauguration of the speaker's corner of our Campus. Paradoxically the most audacious one, due to the importance of the event and the preliminary state of the project.

Tartalo and Galtxagorri were brought out and acted outdoor. No significant body language was shown, neither chatting was possible. Robots were mainly teleoperated and control software was Player/Stage<sup>2</sup>. Only the automatic verse generation system was embedded in wheeled robots.

- **2013/05: Robots against bachelor students:** The robots took part in an event hold in our faculty where they competed against some bertso-amateur students.

Tartalo was accompanied by NAO for the first time. Tartalo acted as troubadour, while NAO acted as the emcee semi-autonomously. Primary gestures were shown by NAO, that was controlled using *Choregraphe*, its native controller.

- **2014/03: Women's day at the Faculty:** Our university annually celebrates the woman's international day in a different centre and in 2014 it was held at our Faculty. The program included a *bertso* event where two big professionals and two robots (NAO and Tartalo) took part.

NAO showed improved chatting abilities, but still "unROSified". Primary gestures were shown in NAO, that guided the event but semi-autonomously.

- **2014/04: Badu, Bada exhibition:** Badu, Bada exhibition offers an area to reflect on and debate about the survival of languages like Euskara, the multilingual world and coexistence between languages. Astigarraga and Rodriguez were invited to give a talk in Bilbao's Alhondiga about *bertsolaritza* and robots, entitled "Minstrel robots: a science or fiction".

This was the first time that NAO acted as verse-maker, and its first performance after being "ROSified". NAO robot showed its verse improvisation and speech-based communication capabilities. The robot only gesticulates while thinking the verse.

- **2014/11: ScienceClub:** Club of Sciences events aim to disclose science and technologies to the society. A dialogue with NAO entitled "Chatting with NAO" of approximately 10 minutes was presented.

---

<sup>2</sup><http://playerstage.sourceforge.net/>

In that event NAO acted semi-autonomously. Gesture and arm imitation-based teleoperation control was shown in the first part of the event. In the second part, the capabilities showed were the same demonstrated in the previous event.

- **2015/11: ScienceClub:** Next year the title of the event was “NAO, an empathetic or just amusing robot?”

Body gestures were integrated and chatting abilities were shown. The key object recognition was tested together with the face and sound localisation behaviours.

- **2016/02: Event at the Faculty:** A local event was organised at the Faculty in order to be able to evaluate the applause classification and emotional expression modules. This time, there was no emcee neither environmental key objects to be easier for the audience to concentrate in the aspects that needed evaluation.

The *bertsolari* robots in front of a seated audience, sang previously generated verses staging only the phase of thinking and singing the verse. After each sung verse the robot gave an emotional response according to the audience feedback.

- **2016/09: Closing of a Summer University Course:** Our university annually organises several summer courses. BertsoBot was invited to the closing of a course entitled “Educational assessment: unresolved matter”. It was not a *bertso-saio* event but it covered all aspects of the interaction.

It was a short exhibition in which NAO showed its dialogue and body expression capabilities. During the scripted dialogue it also sang a verse, and reacted to the audience applause.

- **2017/06: Astigarraga’s thesis presentation** During Astigarraga’s thesis presentation a short *bertso-saio* was carried out to demonstrate the state of the system at that time. Astigarraga acted as the presenter and NAO as troubadour.

The robot showed singing, chatting, and body expression capabilities and its behaviour was influenced by the audience’s emotional state (perceived through applauses). Regarding body expression while talking, no adaptive capabilities were shown, gestures were chosen from a predefined set of gestures.

Table 9.1 summarises the main differences between the first and last version of the BertsoBot system, and the degree of implementation of the main capabilities.

Some of those events were recorded and are available in RSAIT’s YouTube channel<sup>3</sup>.

---

<sup>3</sup><https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ>

	Galtxagorri/Tartalo (2012)	NAO/Pepper (2017)
<b>Dialogue</b>	Precompiled text	Basic chatting capabilities
<b>Poetry Generation</b>	Only one exercise, rhymes given: the system is given the four rhyming words and it is required to compose the bertso. Sentence-retrieval method	Two different exercises: rhymes given and topic given. Sentence-retrieval and n-gram generation methods. Affective state integrated in the creation process.
<b>Affective Perception</b>	No	Audience applause as feedback
<b>Interaction with the environment</b>	Mainly teleoperated robots	Key object recognition integrated, autonomous navigation in the scenario
<b>Body expression</b>	Basic movements (turn on the base and camera tilt movements)	Gesture sets to represent different states of the performance (waiting, thinking, singing, etc), including emotional response gestures to the audience's applause perceived. Automatically generated gestures based on the sentiment of the speech.

Table 9.1: Comparative table showing the capabilities of the first (2012) and last (2018) version of the BertsoBot system





## Part IV

### Conclusions and further work



# Chapter 10

## Conclusions and further work

### 10.1 Conclusions

This research work focuses on social robotics, an area that is strongly related to but goes beyond HRI and service robots. Social robots are starting to become more common in our society and can benefit us as providers of companionship, comfort, entertainment, etc. Many are the applications or purposes of those robots (some of them have been mentioned in Chapter 4). Social robotics is growing fast, and taking news and social media more and more. But it is worth mentioning that social robots are being manufactured rather rapidly even though their capabilities are not yet mature. Before starting this research work we set out two basic questions related to how the interaction with a social robot should be:

1. How do we, humans, communicate (or operate) with social robots?
2. How do social robots act with us?

In this work we have tried to find answers to those questions. In that vein, the work has been developed in two phases: in the first phase we have focused on exploring from a practical point of view several ways that humans use to communicate with robots in a natural manner. Additionally, in the second phase, we have investigated on how social robots must act with the user.

#### **1. Phase: How do we, humans, communicate with social robots?**

We have developed three natural user interfaces intended to make the interaction with social robots more natural. Those interfaces have been tested by developing two applications of different use: guide robots and a humanoid robot control system for entertainment.

With respect to the guide robots application, a system of heterogeneous robots collaborating as guides in multi-floor environments has been developed, which we have called GidaBot. This system has led us to make our wheeled mobile robots more social; the user can interact with the robot through the GUI we have developed, and

the robot communicates with the user in return by using both the GUI and voice. In addition, the system enables robot communication for cooperative guiding tasks in different floors and allows individual navigation in each floor at the same time.

On the other hand, a NAO robot commanding system has been developed for entertainment purposes. Due to their nature, humanoid robots require a more advanced control interface, they have more DoF. This system aims to humanise the way to operate a robot, allowing humans to interact with robots in a way that goes beyond the classic HRI, thus making the interaction more social. The body motion imitation interface and the speech-based interface developed let the user command the robot through natural means, which usually are involved in social interactions, such as body motion, gestures, and voice (including speech recognition, understanding and talking).

## 2. Phase: How do social robots act with us?

Social robots must show social and affective capabilities during interaction with humans. In this phase we focused on identifying and developing the basic behavioural modules that are needed for this type of robots to be socially believable and trustworthy while acting as social agents. We have presented a framework for socially interactive robots that allows the robot to express (some kind of) emotions and to show a natural human-like body language according to the task to be performed and the environmental conditions.

We conclude that an architecture for social robots should be built by at least these building blocks:

1. *Verbal communication* is required to have a conversation with other agents, understanding their requirements and providing appropriate responses.
2. *Perception of the environment* is essential in order to interact with the environment, and allows to identify and recognise objects and other agents in it. Verbal communication is highly enhanced when interlocutor localisation and recognition is shown.
3. *Non-verbal communication (body language)* is needed to make robots expressive in a human-like way, and socially accepted. The interaction is enhanced when the robot conveys information through facial expressions, body gestures or head movements.
4. *Proprioception (body position awareness)* helps to autonomously adapt the actions to be performed according to the current body configuration of the robot. It increases robustness and is fundamental for the system's autonomy.
5. *Affective loop*: perceiving and showing emotions is essential to convey intention. Closing the affective loop makes the user feel more engaged with the system instead of feeling like a mere observer.

All the above enumerated modules should be of course adapted to the ecological niche of the robot. The context, the task to be performed (that delimits the robot behaviour) and the sensory-motor capabilities of the agent to interact within a concrete environment must be taken into account.

Here, we adopted *bertsolaritza* as a test-bed application. This research work takes the *bertsolari* as a cognitive model and presents the Bertsobot architecture, which endows the robots with the required social capabilities to take part in public performances acting and singing verses with other human troubadours and robots. Consequently, the global behaviour is defined according to the features of the improvised poetry sung, and taking into account the competences that *bertsolari*-s show on stage. Bertsobot's architecture includes all the building blocks aforementioned.

Special attention should be paid to the affective loop, fundamental in social robots. To establish the affective loop between users and robots, the latter ones require an emotion perception system that recognises the user emotions, and also a reasoning and response selection/generation mechanism that chooses the emotional response to display at the cognitive level [120] (see Figure 7.1). These mechanisms have been successfully integrated in Bertsobot's architecture. Our robots are able to perceive the audience's feedback measuring the intensity and duration of the audience's applause by means of the applause detection and classification system. In response the robots change their mood and try to better please the public. The change in the mood is reflected, on the one hand, in the body expression, which is shown through subtle gestures, and on the other hand, in the composition of next verse, which is adapted to the sentiment of the message set in the instructions.

Our work had also a secondary goal: to disseminate to the general public the development state of social robots. That is why the validation of the different stages of the development of our social robots was done in public representations (see Section 9.2). These events gave us the opportunity to draw valuable lessons and became an essential tool for qualitatively measuring the social acceptance of the prototypes being developed. In the same way that robots need a physical body to interact with the environment and to become intelligent, social robots need to socially participate in the real tasks they are being built for in order to improve their sociability.

## 10.2 Autonomy of our social robots

Autonomy is a property that can be used to evaluate the "goodness" of the developed framework, but it is difficult to measure since there is no standard definition for it. It is agreed that it is not a yes/no measure and that the degree of autonomy is what should be somehow measured. There is no doubt that social robots must be highly autonomous, no matter what definition of autonomy we take as ground truth.

Lu [106] proposed an ontology of robot theatre that in our opinion is appropriate to measure the state of the Bertsobot. Lu's ontology is based on the automation level and the required control the robots depend on. As you can see in Figure 10.1, Lu divided the area into nine separated regions, which correspond to nine different

classes of robot actor, and further grouped into four larger categories: *Category 1* refers to playback, *Category 2* to teleoperated, *Category 3* to collaborative and *Category 4* to autonomous acting.

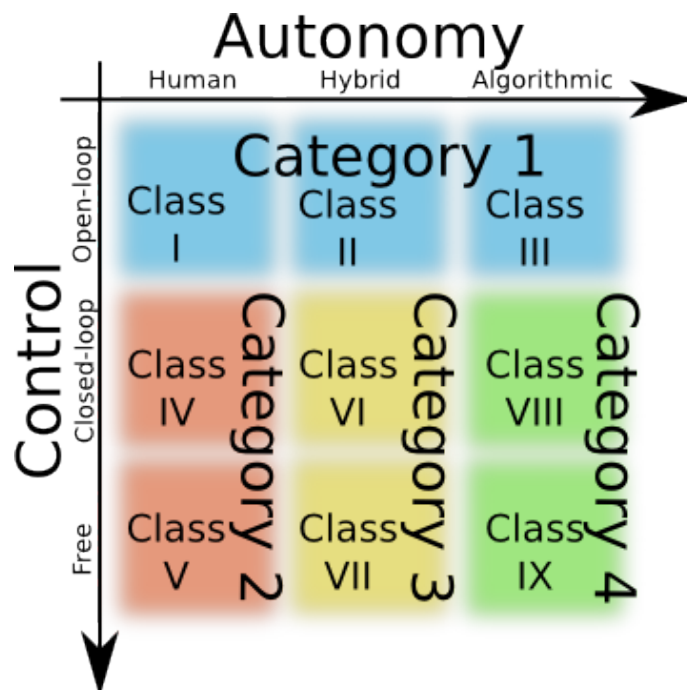


Figure 10.1: Ontology of robot theatre proposed by David Lu

Analysing the evolution of the BertsoBot (see Section 9.2), the first prototype (*Version 1*) utilised in our first performance could be categorised as a Category 1 Class II robot, an open-loop with a hybrid control, hybrid in the sense that its behaviour was partially specified by the human, but there were also algorithmically specified behaviours.

The body motion imitation interface for NAO teleoperation could be classified as Category 2 Class IV, a closed-loop system with human input where the performance changes according to some conditions on the stage but not arbitrarily.

On the other hand, BertsoBot's second prototype (*Version 2*) the robot generates its behaviour via computation, without explicit human intervention further than the oral instructions given by the emcee. The behaviour depends on the robot's own perceptions and the behaviour is produced algorithmically in a closed-loop control. According to those characteristics, it could be categorised as Class VIII. This is a rather forced classification given the fragility of the system. The initial state should always be the same. The lack of self-awareness limits the autonomy.

Regarding the latest prototype (*Version 3*), the affective loop does not increase autonomy itself but improves the degree of sociability and thus, the overall behaviour. The same happens with the capability of spontaneous movement generation. But, it is the body self-awareness behaviour (based on proprioceptive information) which definitively brings the system's autonomy one step forward, and raises the system to the Class VIII without tentative doubt.

Summing up, we are still far from having autonomous free robotic *bertsolari-s* (Class IX in Lu's ontology) but we are little by little making steps forward.

### 10.3 Further work

There are several research lines that we would like to explore more in depth in a near future that we believe might contribute to the development of the area of social robots:

- Emotion appraisal: we have not managed to maintain an internal emotional state, we just modified the basic neutral emotion as a reaction to the current perceptions. To correct this lack we have in mind to define a global emotional state of the robot which will be defined as a combination of the set of internal and external states (see equation 10.1):

$$S = S_{internal} + S_{external} \quad (10.1)$$

Where the internal state is defined by proprioceptive inputs, i.e information obtained from battery level, motor's overheating, etc, and the external state is related to external inputs obtained from the perception of the environment (such as audience's applause, user's face expression and voice intonation) or from the state of the robot after performing the mandated task (whether the task is accomplished or not). All this must be accomplished with a wider set of emotions such as surprise, anger, disgust and fear.

- Emotional behaviour: focusing on how perceived emotions affect the body expression, we intend to investigate the effect of moving the different parts of the body in the same way we have done with the head inclination and the execution velocity of the arm movements. It remains to be analysed how the trunk and arms inclination affects the emotion displayed.
- Self awareness: with respect to the body self-awareness system, this system could also be helpful for any application which requires action planning that depends on the actor's body pose; for instance, to recover the last body position when the robot falls, or to determine which was the last safe position before falling. To that end, we should add fall-down poses into the posture recognition system. Increasing the self-awareness would improve robot autonomy and thus, the global robot's behaviour.
- Learning: social robots need to be adaptive. In the presented work we used learning as a mere tool for implementing concrete behaviours. Generative methods have shown to add a level of spontaneity/naturalness not achieved using deterministic methods. We are aware of the importance of learning for the development of social robots and we consider continuous learning for adaptation a further research area in which we should get involved.

- Intentionality: most robots, the ones we developed included, are merely reactive in the sense that they show no intentionality. They just react as a consequence of a given command. Of course, the intentionality should be explored in the context of the social capabilities the robot is being designed for. This line of research is closely related to that of self decision making. R. Pérula [122] presents a decision-making system based on bio-inspired concepts aimed at deciding the actions to be performed during the interaction between humans and robots that could serve as a starting point for further research.

Finally, we have pending a duel between professional verse-makers and our Bertsobot to show and test the final state of the architecture.



Part V  
Publications



# Chapter 11

## Publications related to Part II

### 11.1 Standardization of a Heterogeneous Robots Society Based on ROS

**Title:** Standardization of a Heterogeneous Robots Society Based on ROS

**Authors:** I. Rodriguez, E. Jauregi, A. Astigarraga, T. Ruiz, E. Lazkano

**Book chapter:** Robot Operating System (ROS)

**Publisher:** Springer

**DOI:** 10.1007/978-3-319-26054-9\_11

**Year:** 2016

# Standardization of a Heterogeneous Robots Society Based on ROS

Igor Rodriguez, Ekaitz Jauregi, Aitzol Astigarraga, Txelo Ruiz  
and Elena Lazkano

**Abstract** In this use case chapter the use of ROS is presented to achieve the standardization of a heterogeneous robots society. So on, several specific packages have been developed. Some case studies have been analyzed using ROS to control particular robots different in nature and morphology in some applications of interest in robotics such as navigation and teleoperation, and results are presented. All the developed work runs for Indigo version of ROS and the open source code is available at RSAIT's github ([github.com/rsait](https://github.com/rsait)). Some videos can be seen at our youtube: channel <https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ>.

**Keywords** Heterogeneous robots · Old robot renewal · Standardization · Teleoperation · Human-robot interaction · Navigation · Speech recognition

## 1 Introduction

The Robotics and Autonomous Systems Lab (RSAIT) is a small research group that focuses its research on applying new machine learning techniques to robots.

The group was founded around year 2000 and inherited a B21 robot (RWI). Since then, the group has grown up and, in its development, has acquired different robots.

---

I. Rodriguez (✉) · E. Jauregi · A. Astigarraga · T. Ruiz · E. Lazkano  
Faculty of Informatics, Robotics and Autonomous Systems Lab (RSAIT),  
UPV/EHU, Manuel Lardizabal 1, 20018 Donostia, Spain  
e-mail: [igor.rodriiguez@ehu.eus](mailto:igor.rodriiguez@ehu.eus)  
URL: <http://www.sc.ehu.es/ccwrobot>

E. Jauregi  
e-mail: [ekaitz.jauregi@ehu.eus](mailto:ekaitz.jauregi@ehu.eus)

A. Astigarraga  
e-mail: [aitzolete@gmail.com](mailto:aitzolete@gmail.com)

T. Ruiz  
e-mail: [txelo.ruiz@ehu.eus](mailto:txelo.ruiz@ehu.eus)

E. Lazkano  
e-mail: [e.lazkano@ehu.eus](mailto:e.lazkano@ehu.eus)

In our experience, robot maintenance is laborious, very time consuming, and does not provide immediate research results. Moreover, robots from different suppliers have their own control software and programming framework that require senior and incoming members to be trained once and again. This makes robot maintenance harder. That's why robotics labs often become scrap yards in the sense that old robots are often retired instead of upgraded and broken robots are discarded instead of repaired. But small research groups often do not have enough budget to invest in new robots, so that upgrading and repairing robots become mandatory.

We now own a heterogeneous set of robots, consisting of an old B21 model from RWI named *MariSorgin*; a *Kbot-I* from Neobotix; *Galtxagorri* a Pioneer 3DX and the PeopleBot *Tartalo*, both from MobileRobots; a humanoid NAO from Aldebaran; five *Robotino*-s from Festo (these ones used for educational purposes in the Faculty of Informatics). Each one came with its own API and software, most running on Linux. Thanks to ROS we now have a standard tool to uniformly use this society of heterogeneous robots.

**Contributions of the book chapter:** several case studies are presented in which some new ROS drivers and packages have been developed for navigation and gesture and speech based teleoperation that can be used for robots different in nature. Those applications are fully operative in real environments.

## 2 Robot Description

A brief description of the robots and the modifications and upgrades suffered during their operational life follows up, together with a reference to the software used to control them.

### 2.1 *MariSorgin*

Our heirloom robot is a synchro-drive robot that dates from 1996. It is a B21 model from Real World Interface provided with a ring of ultrasound, infrared and tactile sensors for obstacle avoidance. Opposite to its successor, the well known B21r model, it was not supplied with a laser sensor. In 2002 its motor controllers were damaged and sent to RWI for replacement, but they never came back to us. Ten years later, those boards were replaced with Mercury motor controllers from Ingenia Motion Control Solutions [8]. The two internal i386 PCs were replaced with a single newer motherboard. So, after 10 years *MariSorgin* became again fully operational.

In the beginning the original API, named *BeeSoft* [19] was replaced by a home made library that better suited to our control architecture development philosophy (*libB21*<sup>1</sup>). We combined it with *Sorgin* [2], a framework designed for developing

---

<sup>1</sup>Developed by I. Rañó at Miramon Technology Park, 2001 [17].

```

void main()
{
    /* Data declarations */
    io_data_t laser_readings;
    io_data_t motor_output;

    /* Behavior declaration */
    behavior_t avoid_obstacles;

    /* Data initialization */
    io_data_alloc(&laser_readings, 181, NULL, 0, 0);
    io_data_alloc(&motor_output, 2, NULL, 0, 0);

    /* Behavior initialization */
    behavior_define(2, i, avoid_obstacles_start,
                  avoid_obstacles_stop,
                  avoid_obstacles_calculate);

    /* input/output connections */
    behavior_set_input(&avoid_obstacles, 0, running);
    behavior_set_input(&avoid_obstacles, 1, laser_readings);
    behavior_set_output(&avoid_obstacles, 0, motor_output);

    behavior_start(&avoid_obstacles);
    behavior_run(&avoid_obstacles);

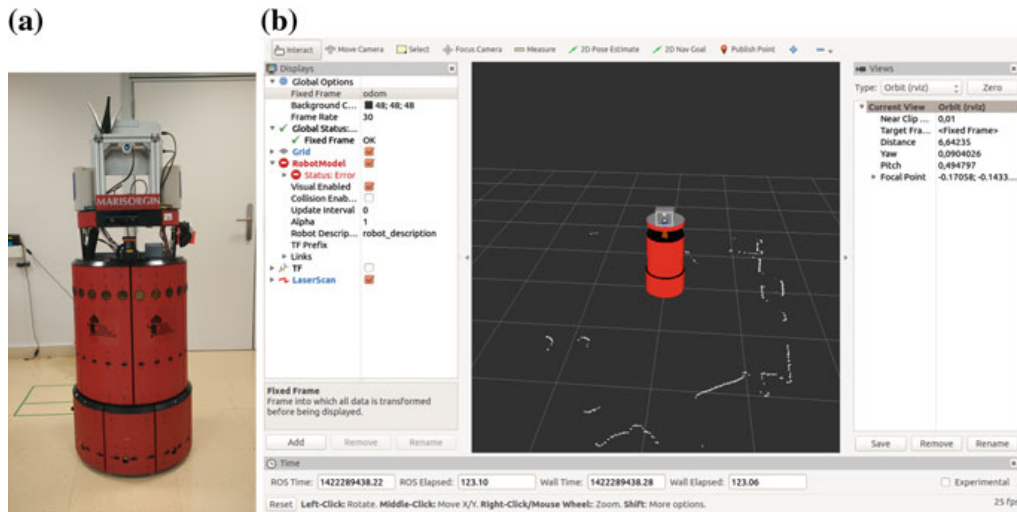
    sleep(100);
    behavior_stop(&avoid_obstacles);
}

```

**Fig. 1** *Sorgin*: example program

behavior-based control architectures. *Sorgin* allowed us to define and communicate behaviors in a way similar to ROS topics and nodes. Topics equivalents were arrays of floats defined as `io_data` structures, and nodes were `behavior_t` structures, with different associated functions (initialization, main loop and stop) launched in separated threads. Thus, *Sorgin*'s modular structure resembled ROS procedural organization and communication but in a more modest implementation. Figure 1 shows what a *Sorgin* program looks like.

After the “resurrection”, we had no doubt: we must adapt it to ROS. So, we developed the necessary ROS drivers for the motor controllers and mounted a Hokuyo URG-30 laser on top of the enclosure, a Kinect camera and a Heimann thermal sensor (see Fig. 2).



**Fig. 2** *MariSorgin* and its URDF model. **a** Renewed B21, **b** URDF model visualized in Rviz

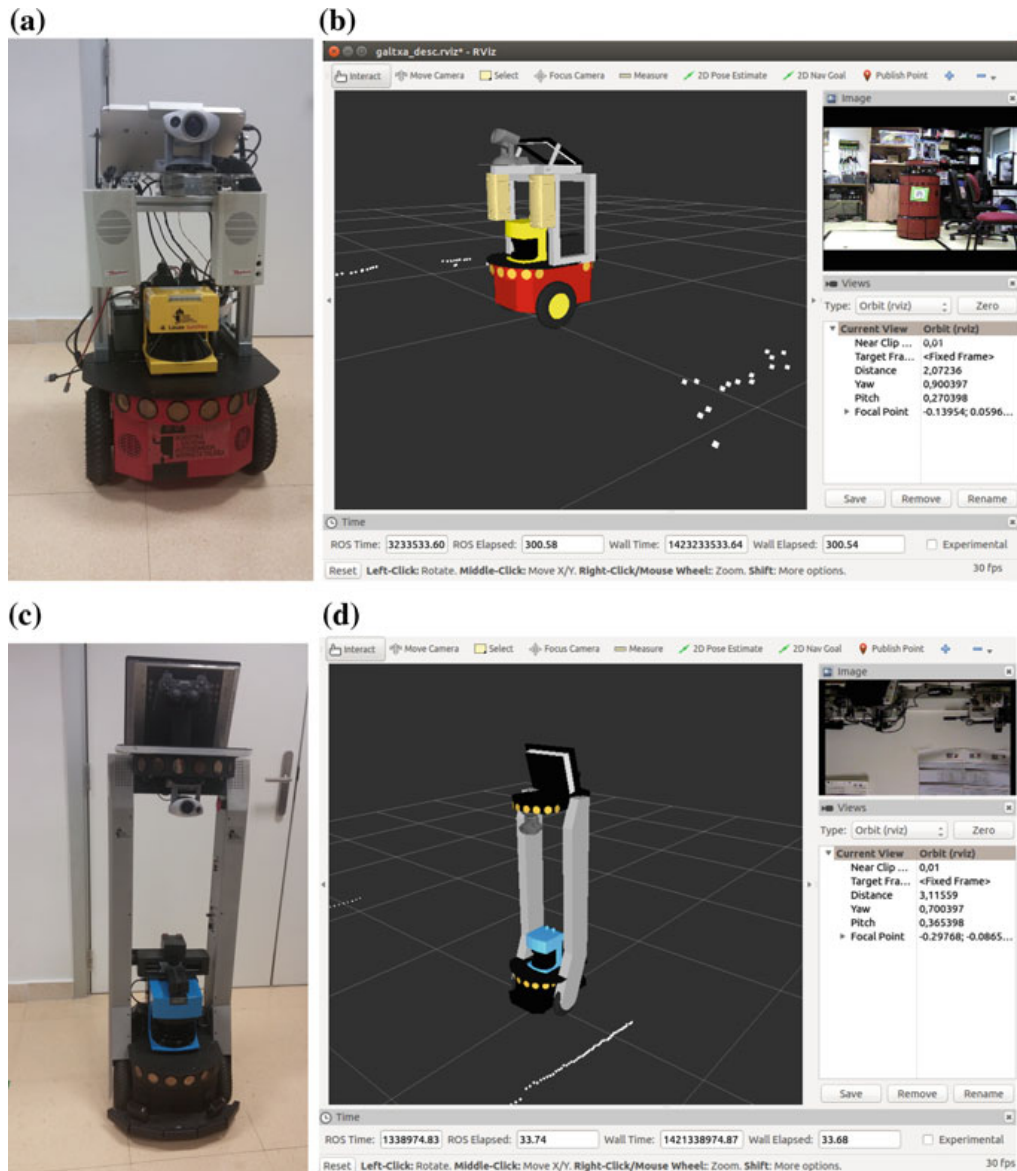
## 2.2 *Tartalo and Galtxagorri*

Two differential drive robots from MobileRobots. *Galtxagorri*, a Pioneer-3DX robot suffered some modifications from its initial configuration. On the one hand, a Leuze-RS4 laser sensor was mounted on top of its body (fed externally to extend the duration of the internal batteries and so, the robot's autonomy). Speakers have been added together with an amplifier (Fig. 3). Besides, *Tartalo* is a PeopleBot robot that facilitates human-robot interaction. Both platforms have a MAMBA VL-EBX-37A board with a 2.26 GHz Intel(R) Core(TM2) Duo CPU.

These two robots came with *Aria*, a framework that again did not fulfill our control architecture development schemata. Before ROS came up, our trend was to use Player/Stage (see [27]). Player/Stage offered us a wide set of drivers and a proper tool for developing our own algorithms without imposing restrictions in the type of control architecture being developed.

Player/stage shares with ROS the definition of what a robot is, i.e. a set of devices (sensors and actuators), each one with its own driver that gives access to the device data. Player offered an abstract layer of interfaces that allowed to access different devices similar in nature using the same code. Combining Player with *Sorgin* turned out to be straight forward (Fig. 4). This coupling allowed us to work with MobileRobots platforms in a flexible and suitable way for several years. But Player focused more on developing drivers than algorithms and stopped evolving when ROS appeared.

ROS provides the *P2OS* package that allows to control *Tartalo* and *Galtxagorri*'s base. Moreover, it also offers the appropriate driver for the Leuze RS4 laser scanner. Thus, only the URDF models were needed to set up for the two robots.



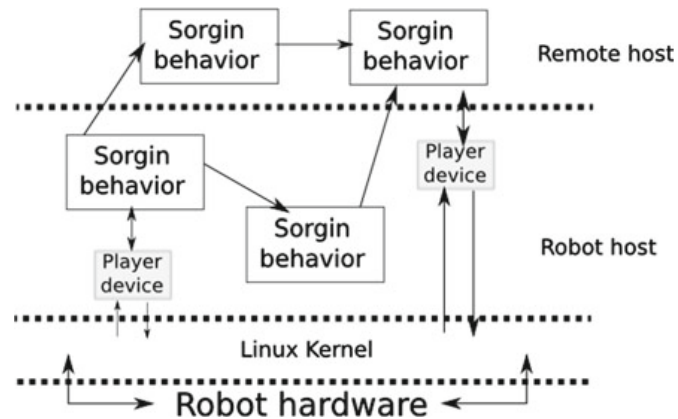
**Fig. 3** P2OS robots and their URDF models. **a** Galtxagorri. **b** Galtxagorri in Rviz. **c** Tartalo. **d** Tartalo in Rviz

### 2.3 Robotino-s

Those are omnidirectional circular platforms from Festo Didactic that we mainly use for education. They are provided with several sharp GP2D12 infrared sensors, a bumper ring, a webcam and a Hokuyo URG-04LX sensor in order to be able to experiment with mapping and planning techniques. The control unit, placed on top of the wheeled platform, contains a 500 MHz PC104 processor that runs RTLinux.



**Fig. 4** *Sorgin+Player*: organization



All the software is installed in a 4 GB Compact Flash card. The unit also offers an Ethernet port and a WLAN access point, 2 USB ports and a VGA connector.

*RobotinoView* is the interactive, graphic programming environment for Robotino. Besides, Robotino ships with an application programming interface (*RobotinoAPI2*) allowing the user to create programs using different programming languages like C, C++, Java and more. Communication between the control program and Robotino is handled via TCP and UDP and is therefore fully network transparent. The new API2 is based on a RPC like infrastructure. The REC-RPC library is an interprocess communication middleware similar to ROS. It is completely based on Qt and does not have any other dependencies.

Migration to ROS has been straight forward, since packages for *Robotino* can be found at [wiki.ros.org/robotino](http://wiki.ros.org/robotino).

## 2.4 NAO

NAO is an autonomous programmable humanoid robot developed by Aldebaran Robotics. NAO's human like shaped body is about 58 cm tall and weights about 4,8 kg. It is built in polycarbonate and ABS (a common thermoplastic) materials that allow better resistance against falls and it has a lithium battery with which it can get an autonomy of 90 min approximately. Its heart is composed by a 1.6 GHz Intel Atom processor running Linux. 25 servos enable to control the 25° of freedom of the robot. Regarding to robot motion, NAO can move in any direction (omnidirectional walking), it uses a simple dynamic model (linear inverse pendulum) and quadratic programming. It is stabilized using feedback from joint sensors. It can walk on a variety of floor surfaces, such as tiled and wooden floors, and he can transition between surfaces while walking.

NAO sees using two 920p cameras, which can capture up to 30 images per second. Also, it uses four microphones to track sounds and two loudspeakers to talk or play sounds.

*Choregraphe* is the original programming software of NAO. It is a multi platform desktop application that allows to create animations and behaviors, test them on a simulated robot, or directly on a real one and monitor and control the robot. *Choregraphe* allows to create very complex behaviors (e.g. interaction with people, dance, send e-mails, etc.) without writing a single line of code. In addition, it allows the user to add her own Python code to a *Choregraphe* behavior.

The behaviors created with *Choregraphe* are written in its specific graphical language that is linked to the *NAOqi* Framework, the main software that runs on the robot. NAO interprets them through this framework and executes them. *Choregraphe* also interacts with *NAOqi* to provide useful tools such as the Video monitor panel, the Behavior manager panel, the Toolbar, the Robot view or the Timeline Editor.

Again, transition to ROS was easy. ROS drivers for NAO can be found at <http://wiki.ros.org/nao>.

## 2.5 *Kbot-I*

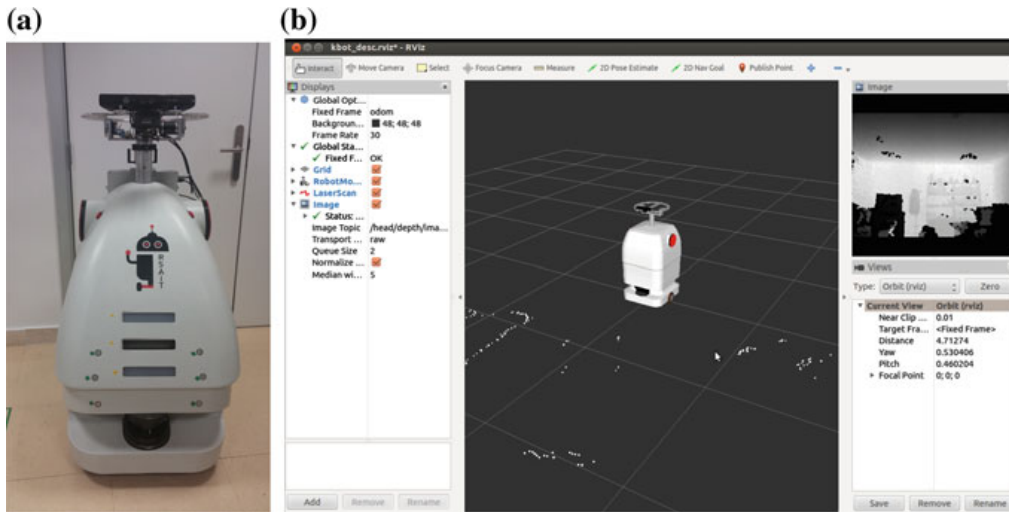
A differential drive robot built by Neobotix in 2004 for acting as a tour guide at the Eureka Museum of Science in San Sebastian. Supplied with a Sick S3000 laser scanner, the robot held a touch screen for receiving orders. An application specifically developed for the robot to act as a guide in the museum was running on the onboard Windows 2000 Professional machine. In 2006 the robot was damaged and stored in a garage until 2014. The robot was transferred to the University of the Basque Country (UPV/EHU). In spite of the lack of any detailed manual, our group managed to locate and repair broken connections. The onboard PC was replaced by a Zotax MiniPC with a NVIDIA graphics card, and the webcam on its head was removed and instead, a Kinect sensor has been mounted. The rigid arms that supported a huge touch monitor were also removed and replaced by plates built with a 3D printer (another invaluable tool for robot maintenance!). A smaller Getich monitor has been placed on the back side of the body. Fortunately, the source code of the drivers was available and only small modifications were needed to compile that code and make the necessary libraries under Linux. Albeit the time spent in code surfing, it was rather straightforward to implement the necessary ROS drivers to get it back running<sup>2</sup> (Fig. 5).

## 3 Working Areas of RSAIT Research Group

Navigation is a fundamental skill that mobile robots need in order to be autonomous. The navigation task has been approached in different ways by the main paradigms of control architectures. RSAIT has focused its navigation methodology within the

---

<sup>2</sup>Thanks to Marco Beesk from Neobotix for agreeing to make public our *Kbot-I*ROS nodes.



**Fig. 5** *Kbot-I* and its URDF model. **a** Renewed *Kbot-I*, **b** *Kbot-I* visualized in Rviz

behavior-based philosophy (see for instance [9]) that focuses on biology to inspire its navigation strategies ([12]). But probabilistic approaches seem to increase adepts and some techniques are being distributed within the ROS community for mapping, localization and planning. No definitive solution exists nowadays but clearly, ROS navigation stack makes possible to compare different approaches. Therefore, it is worth to setup this stack for our robots.

Besides, in our research group we are working on different applications for natural human-robot interaction. On the one hand, we have developed two different ROS packages to enrich the teleoperation of robots: speech-based teleoperation in Basque Language (*Euskara*) and gesture-based teleoperation using the Kinect [18].

On the other hand, we have developed a system, called **Bertsobot**, which is able to construct improvised verses in Basque (named *bertsoak*) according to given constraints on rhyme and meter, and to perform them in public (see [1]). NAO is the robot that gives shape to the **Bertsobot** system. It is capable of understanding some “orders”, composing and playing traditional Basque impromptu verses, also replicating the movements made by the impromptu verses singers. This project allowed us to combine diverse research areas such as body gesture expressiveness, oral communication and human-robot interaction in a single project.

Table 1 summarizes the developed ROS packages. The experiments described in the following sections will explain how these skills have been integrated in the different robots, according to their sensorial capabilities.

**Table 1** Summary of basic ROS modules

	Used ROS modules	Adapted ROS modules	New ROS modules
General use			-Speech_eus: Basque TTS and ASR modules -heiman: thermopile driver
<i>Galtzagorri</i>	-p2os_driver -rotoscan_node -gscam	-galtxa_description: URDF model	-galtxa_teleop_speech_eus
<i>Tartalo</i>	-p2os_driver -sicklms -gscam	-tartalo_description: URDF model -tartalo_navigation: planner and costmap params	-tartalo_teleop_speech_eus
<i>Robotino</i>	-robotino_node -openni_node -openni_launch	- All packages catkinized - skeleton_tracker: openni_tracker modified	-robotino_teleop_gestures
<i>Kbot-I</i>	-sicks300 -openni_node -openni_launch		-kbot_description: URDF model -kbot_platform: drivers for driving motors, head tilt motor and integrated sensors (US) -kbot_teleop_joy: platform and head tilt motor control -kbot_guideqt: interactive user interface for navigation
<i>MariSorgin</i>	-hokuyo_node -imu_um6 -openni_node -openni_launch		-mari_description: URDF model -cann: mercury motor controller driver -mari_teleop_joy: platform control -mari_teleop_speech_eus -mariqt: user interface for speech based teleoperation -heiman: thermopile driver

(continued)

**Table 1** (continued)

	Used ROS modules	Adapted ROS modules	New ROS modules
NAO	<ul style="list-style-type: none"> <li>- naoqi_bridge</li> <li>- nao_robot</li> <li>- nao_meshes</li> <li>- nao_interaction</li> <li>- nao_extras</li> </ul>	<ul style="list-style-type: none"> <li>- skeleton_tracker: openni_tracker modified</li> </ul>	<ul style="list-style-type: none"> <li>- nao_teleop_gestures: motion, hands and head controller</li> <li>- nao_teleop_speech_eus</li> <li>- nao_bertsoBOT</li> </ul>

## 4 Case Study 1: Setup of the Navigation Stack

*Navigation* refers to the way a robot finds its way in the environment [13]. Facing this is essential for its survival. Without such a basic ability the robot would not be able to avoid dangerous obstacles, reach energy sources or return home after exploring its environment. Navigation is therefore a basic competence that all mobile robots must be equipped with. Hybrid architectures tackle the problem of navigation in three steps: mapping, localisation and planning. These are old problems from the perspective of manipulation robotics and are nowadays treated in a probabilistic manner. Hence the name of the field *probabilistic robotics* [26], that makes explicit the uncertainty in sensor measurements and robot motion by using probabilistic methods. ROS offers several stacks that use probabilistic navigation techniques and allow to empirically use, test and evaluate the adaptability of those techniques to different robot/environment systems. So far, and taking as starting point the navigation stack available for the P2OS robots, it has been setup in *Kbot-I*.

Since *Kbot-I* is now ready again for human-robot interaction, an interactive user interface has been developed using **rqt** (`kbot_guide_qt`) to retake the original task *Kbot-I* was designed for: be a guide within our faculty. The most frequently demanded sites of our faculty are located at the first floor. Hence, in this attempt a map of the first floor has been created with ROS mapping utilities and that map is being used as the floor plan of the developed GUI. This floor plan has been populated with several interaction buttons corresponding to the important locations people might be interested in, such as the administration, the dean's office, the lift, several labs and so on. Information about actual and destination locations is also displayed on the interface. Figure 6 shows what the GUI looks like.

The robot morphology makes door crossing insecure and thus, for the time being the GUI limits the robot guiding task to the front of the door that gives access to the desired location.

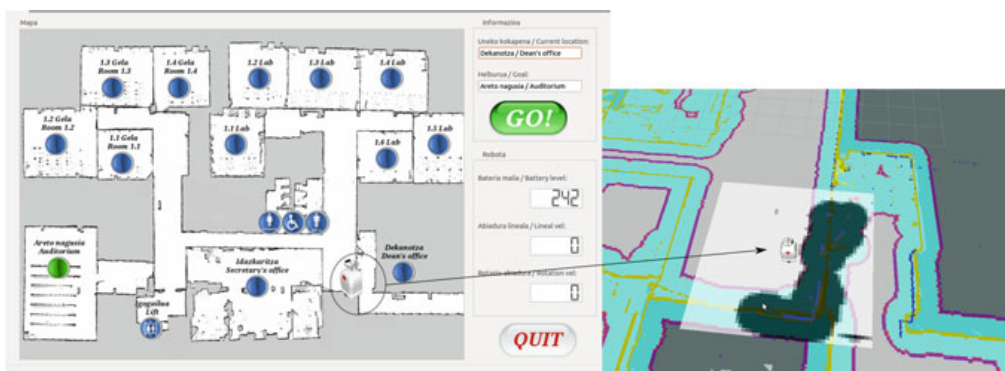


Fig. 6 Kbot navigation interaction window and costmap

*System evaluation:* Regarding the setup of the navigation stack, it is not a friendly process. Although the documentation has been improved, many parameters have to be set empirically, without any explicit methodology. The ROS navigation stack is based on probabilistic navigation techniques and it is known that they don't adapt well to dynamic environments. The system fails when there are severe mismatches between the current sensor readings and the stored map. Therefore, crowds should be avoided in front of the robot during the tours and all people must be advised to stay on the back of the robot so that the map remains reliable and the planner could find a way to the goal. Moreover, the application needs to know the robot's initial position in order to be able to plan routes to the goal. Hence, it is not able to face the global localization problem. It will be a great improvement to enhance the navigation stack with global localization capabilities to overcome this problem and to make it more robust and general to use.

But more important is to mention that, after ROSifying the robot, we got a navigation application running, working and prepared to be used in public in just a couple of weeks, but without the need of reimplementing the whole system. The application has been used for the first time in an open door event at our faculty on March 12 (2015). About 100 candidate students came to visit the faculty and they were divided on 6 small groups of 15–20 students. They were supposed to visit different labs and sites on different floors of the building. The robot was located on the first floor and guided the teams over the different places they should arrive to. Basque TV (EiTB) came to record the event and broadcasted it at the news (Fig. 7).



**Fig. 7** Kbot making guided tours in the faculty



Still we can improve the system integrating door crossing abilities. Also, the system should be complemented with the maps of the second and third floors. Our plan is to set *Tartalo* in the second floor and *MariSorgin* in the third one, so that connection among floors will be done via the lift. Robots will not entry the lift but will communicate to be aware that they need to welcome “tourists” sent from other locations.

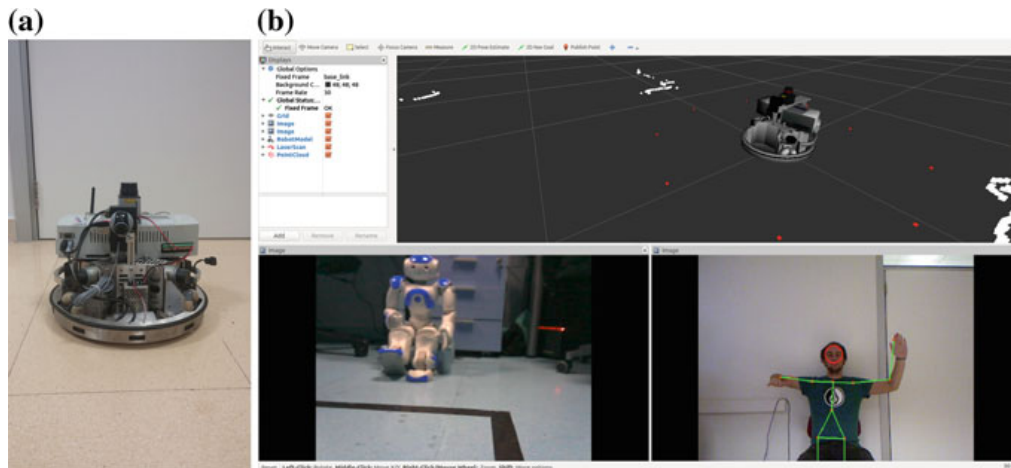
## 5 Case Study 2: Kinect Based Teleoperation

The term teleoperation is used in research and technical communities for referring to operation at a distance. Teleoperated robots are used in many sectors of society. Although those robots are not autonomous they are very useful e.g. in medicine for surgery [4, 5], for space exploration [3] or for inspection in nuclear power plants [16].

Different devices can be used for teleoperating a robot (joystick, smart phone, wii-mote) but gesture based teleoperation is increasing adepts ([6, 15, 25]) specially due to availability of cheap 3D cameras such as Microsoft’s Kinect sensor. Real-time teleoperation of humanoid robots by detecting and tracking human motion is an active research area. This type of teleoperation can be considered as a particular way of interaction between a person and a robot, because it is a natural way to interact with robots. It is an interesting research topic and related work is abundant. For instance, Setapen et al. [21] use motion capture to teleoperate a NAO humanoid robot, using inverse kinematic calculations for finding the mapping between motion capture data and robot actuator commands. Matsui et al. [14] use motion capture to measure the motion of both, a humanoid robot and a human, and then adjust the robot motion to minimise the differences, with the aim of creating more naturalistic movements on the robot. Song et al. [22] use a custom-built wearable motion capture system, consisting of flex sensors and photo detectors. To convert motion capture data to joint angles, an approximation model is developed by curve fitting of 3rd order polynomials. Koenemann and Bennewitz [10] present a system that enables a humanoid robot to imitate complex whole-body motions of humans in real time, ensuring static stability when the motions are executed and capturing the human data with an Xsens MVN motion capture system consisting of inertial sensors attached to the body.

The above mentioned methods are limited in the sense that the human needs to wear different types of sensors in order to interact with the robot. This can be avoided with the Kinect sensor, moreover, the cost of the equipment is declined. That is why researchers have become more interested in Kinect. Song et al. [23] propose a teleoperation humanoid robot control system using a Kinect sensor to capture human motion and control the actions of remote robot in real-time. Suay and Chernova [24] present a new humanoid robot control and interaction interface that uses depth images and skeletal tracking software to control the navigation, gaze and arm gestures of a humanoid robot.





**Fig. 8** *Robotino* and its teleoperation interface. **a** *Robotino*, **b** The teleoperation interface

ROS offers drivers for the Kinect together with a package that extracts and tracks the human skeleton from sensor data. Thus, taking as base tool these two packages (*openni\_launch* and *openni\_tracker*), a gesture-based teleoperation system has been developed for a holonomic wheeled robot and, afterwards, enriched to teleoperate a humanoid robot.

### 5.1 *The robotino\_teleop\_gesture Package*

The development of a gesture based teleoperation system requires first to identify the degrees of freedom that are going to be controlled and define the set of gestures that will control the robot. *Robotino-s* are holonomic wheeled robots and thus, can be moved along the plane in any direction without changing the robot heading. Also, a rotational velocity can be assigned. The defined gesture set is based on arm movements although internally is implemented through hand positioning (see Fig. 8). The gesture set consists of:

- Moving right arm tilt controls forward/backward movements (lineal velocity in x)
- Right arm pan movement controls side movements (lineal velocity in y)
- Left arm yaw movement controls left/right rotation (rotational velocity)
- Lowering both arms at the same time stops the robot.

The developed teleoperation system has two sides: the user detection process and the robot motion control process. On the one hand, the user detection step is based on the *openni\_tracker* package, but several changes have been introduced to produce an *skeleton\_tracker*:

1. The node that tracks the skeleton now publishes the joint position information of the skeleton in the `skeleton` topic.
2. A new node makes available the Kinect image that includes the graphical representation of the skeleton links on it.

On the other hand, the `robotino_teleop_gesture` node contains a subscriber that receives messages published by `skeleton_tracker` in the `skeleton` topic. When a message is received, the operator's position is analyzed. And according to that position the robot executes the corresponding motion. Although each arm movement controls a velocity value, different gestures can be combined, i.e. move forward while turning.

*System evaluation:* The skeleton tracker performs properly when the only moving element of the scene is the teleoperator and so, the background needs to be static. Moreover, the system setup is designed for a single person sat on a chair in front of the kinect. But when the application is used with children, they must stand up so that size does not affect the calibration process. The application is fully operational for real indoor environments and is being used as a game/demo in several yearly events like the week of sciences (2013–2014), meetings with undergraduate students (2012–2015), robotics day (2013). Since its early development, it has been adapted for several ROS distros and *OpenRobotinoAPI* versions. Up to now, it is catkinized for Indigo and *OpenRobotinoAPI* version 0.9.13.

## 5.2 *The nao\_teleop\_gesture Package*

The gesture-based teleoperation system developed for the *Robotino-s* has been adapted and extended to be used with NAO. The skeleton tracking system is exactly the same, the only difference is that more degrees of freedom are to be controlled and, thus, the gesture set needs to be redefined and extended.

NAO's human like morphology allows not only the motion of the robot in the plane but also the movement of the arms. Thus, it is not adequate to use the operator arms to control the velocities of the robot. In this case, the selected gesture set is the following:

- If the operator steps forward/backward the robot walks forward/backwards.
- The lateral steps of the operator cause the side movements of the robot.
- Raising the left shoulder and lowering the right one causes clockwise rotation.
- Raising the right shoulder and lowering the left one causes ccw rotation.
- Left/right arm movements are used to control robot's left/right arm.
- Head pan and tilt movements are used to control NAO's head.

The new package, named `nao_teleop_gesture` contains three nodes:  
 1. - `nao_motion_control`: basically, this node has the same functionality as the node developed for the *Robotino-s*. It has to perform the following two main tasks:

- Receive messages published by the `skeleton_tracker` package.
- Publish NAO's walking velocities.

The `nao_motion_control` node has a publisher that publishes the omnidirectional velocity ( $x$ ,  $y$ , and  $\theta$ ) in the `cmd_vel` topic for the walking engine. The velocity with which the robot moves has been set to a constant value. If the walking velocity is too high the robot starts to swing. Thus, the linear velocity and the angular velocities have been assigned low values.

2. - `nao_arm_control`: This node is in charge of sending to the robot the necessary motion commands to replicate the operator's arms motion. The node performs tasks by:

- Receiving the messages published by the `skeleton_tracker` package.
- Publishing NAO's joint angles with speed.

Therefore, `nao_arm_control` is subscribed to the `skeleton` topic in order to receive the operator's skeleton messages published by `skeleton_tracker`.

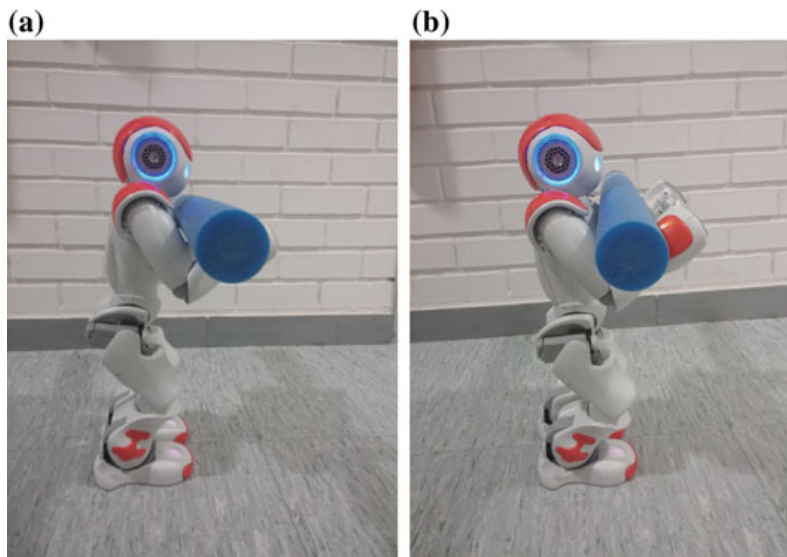
On the other hand, the `nao_arm_control` node has a publisher that publishes the joint angles with speed in the `joint_angles` topic, which allows the communication with the `nao_controller` node. The NAO's joints motion speed is set to a constant value appropriate for the robot to mimic the operator arms motion in "real" time.

3. - `nao_head_control`: This node is responsible of moving the robot's head. Similar to the way that `nao_arm_control` gets the arm joint angles, this node calculates the head joint pitch and yaw angles, and publishes them into the `joint_angles` topic.

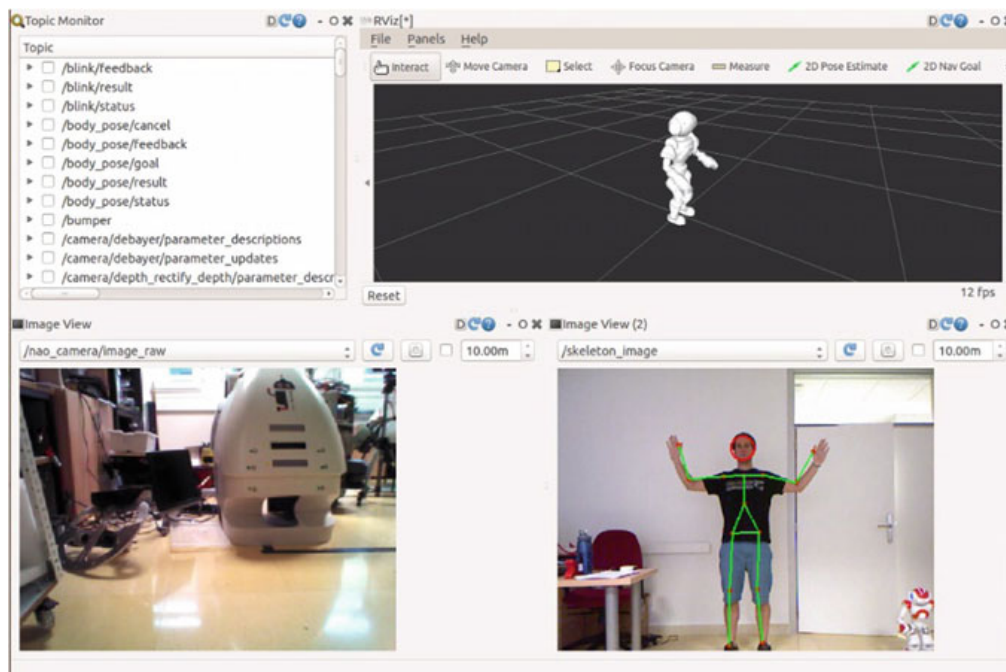
The robot imitates human actions in real-time with a slight delay of less than 30 ms. This delay is approximately the time the system needs to capture the operator's arms/head motion, calculate the angles that make up the operator's arms/head joints (see [18]), and send motion commands to the robot via WiFi.

Only walking action movements (forward, backward, left, right) with rotational motions can be combined. No arm movement is allowed while walking so that the stability of the robot is not affected. Moreover, when the robot holds something on its arms the center of gravity (COG) of the walking robot needs to be lowered and backwarded, so that the COG is maintained within the support polygon (see Fig. 9). Thus the walking behavior has been modified for those cases in order to increase the stability.

A GUI has been created (this time with existing *rqt* plugins) in order to help the operator to know the system state. The GUI is divided into two main parts (Fig. 10). The top side is composed by the *Topic Monitor* and the *Rviz* interface. The *Topic Monitor* shows all the topics and messages sent by the nodes that are in execution. *Rviz* shows the NAO 3D model moving in real-time. The bottom side shows visual information from the cameras; *Image View* shows the image received from NAO's top camera and the right window shows the image captured by the Kinect together with the skeleton of the tracked body.



**Fig. 9** Modified walking position. **a** Original. **b** Modified



**Fig. 10** Teleoperation display

The system starts with NAO in crouching position and when the operator enters the Kinect's view, the calibration process starts. NAO tells the operator that the calibration ended successfully saying "Kinect control enabled" and then, the operator can control the robot with his/her body.

*System evaluation:* Imitation is an important way of skill transfer in biological agents. Many animals imitate their parents in order to learn how to survive. It is also a way of social interaction. A sociable robot must have the capability to imitate the agents around it. In a human society, people generally teach new skills to other people by demonstration. We do not learn to dance by programming, instead we see other dancers and try to imitate them. Hence, our artificial partners should be able to learn from us by watching what we do. That idea pushed us to evaluate our application based on the imitation ability of the robot.

Two experiments were defined to evaluate the system. Those experiments involved several people that should give qualitative measures of the system performance by means of a questionnaire that participants completed after carrying out each experiment. Experiments were performed until each participant achieved the aim of the experiment at least once (see [18]). The experiments revealed three aspects that might be improved:

- The lack of side view makes more difficult the guidance of the robot. This problem is now alleviated with the addition of the head motion control.
- Although the selection of gestures is correct (natural) and the movements are quite precise, a short period of training is needed by the operator to get used to distances.
- The robot can lose balance when walking with the arms raised.

## 6 Case Study 3: Speech Based Teleoperation in Basque

Human-robot interaction (HRI) is the study of interactions between humans and robots. HRI is a multidisciplinary field with contributions from human-computer interaction, Artificial Intelligence, robotics, natural language understanding, design, and social sciences. A requirement for natural HRI is to endow the robot with the ability to capture, process and understand human requests accurately and robustly. Therefore it is important to analyse the natural ways by which a human can interact and communicate with a robot.

Verbal communication should be a natural way of human-robot interaction. It is a type of communication that allows the exchange of information with the robot.

To serve a human being, it is necessary to develop an active auditory perception system for the robot that can execute various tasks in everyday environments obeying spoken orders given by a human and answering accordingly. Several systems have been recently developed that permit natural-language human-robot interaction. Foster et al. [7] propose a human-robot dialogue system for the robot JAST, where the user and the robot work together to assemble wooden construction toys on a common workspace, coordinating their actions through speech, gestures, and facial displays.

A speech based teleoperation interface should provide the user the possibility to teleoperate the robot giving predefined orders [28]. The system also should give feedback to the operator when an instruction is not understood and this feedback should also be verbal. Three elements are identified in an architecture for speech-based teleoperation:

1. The automatic speech recognition system (ASR)
2. The text to speech (TTS) system
3. The robot control system

The first two elements are robot independent and, thus, have been integrated in a single package named **speech\_eus**. This package contains two nodes, one responsible of the speech recognition step and the second one, responsible of the text to speech translation. Let's describe the nodes that compose the `speech_eus` package.

1. *gspeech\_eus* node: Our robots are supposed to interact in **Euskara** (Basque, a minority language spoken in the Basque Country) and thus, a tool adapted to this requirement was needed. ROS *gspeech* package gives ASR capabilities to the robot and can be configured for many languages, including Basque. But this package needs some modifications in order to be useful in a real-time teleoperation scenario. These are the introduced changes:

- When the native `gspeech` runs the Google Speech Service, it is executed only once, i.e. when the user starts speaking, the audio is captured and sent to Google. There it is analysed and the text “corresponding” to the received audio is returned with a confidence level; then, the program ends. It could be tedious for the user to run the speech recognition node each time she/he wants to order something to the robot, or each time she/he receives an error message. Hence, the new node now runs iteratively avoiding the problem of having to launch the node each time the user wants to talk.
- When the Google Speech Service does not recognize the spoken words, it returns an error message and then the node is forced to quit. Now, error messages received from Google Speech Service are specially treated. If an error message is received, `gspeech_eus` publishes a `Repeat` message in the `google_speech` topic to advertise the user that his/her spoken words are not being recognized.
- The original `gspeech` node only prints the response received, it does not publish any messages or services, so it can not communicate with other nodes. After the modifications, the confidence level of the hypothesis received from *Google Speech* is processed and, if it is lower than a predefined threshold (0.15 for the performed experiments), the response is declined and treated as an error message.

2. *tts\_eus* node: This is the node in charge of converting the text into speech using the *AhoTTS* tool [11]. That system, developed by the Aholab group in the University of the Basque Country, is a modular text to speech synthesis system with multithread and multilingual architecture. It has been developed for both, Euskara and Spanish languages. The TTS is structured into two main blocks: the linguistic processing module and the synthesis engine. The first one generates a list of sounds, according to the Basque SAMPA code [20], which consists of the phonetic transcription of the expanded text, together with prosodic information for each sound. The synthesis engine gets this information to produce the appropriate sounds, by selecting units and then concatenating them and post-processing the result to reduce the distortion that





**Fig. 11** Setup for the experiments

appears due to the concatenation process. This tool is required for communicating in Basque Language, but it would not be required for English interlocation.

`tts_eus` has a subscriber that receives messages from the `text_speech` topic. When a text message is received, this node converts it into speech (an audio file) and plays the audio over robot's speakers.

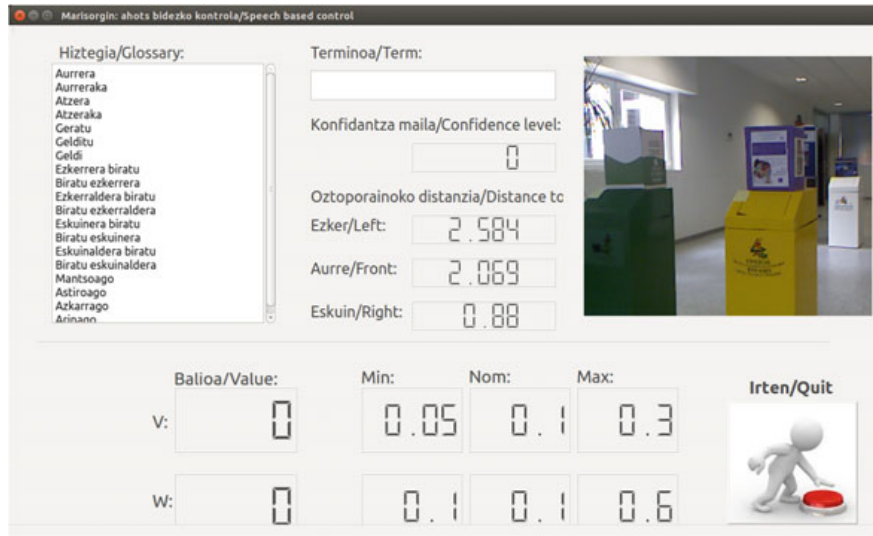
## 6.1 *Speech-Based Teleoperation in MariSorgin*

Again, this teleoperation system has two sides: the instruction interpretation process and the motion controller. Regarding to the instruction interpretation part, and as mentioned before, our robots are supposed to interact in **Euskara**. The oral commands are captured by a microphone and sent to the Google Speech Service by the `gspeech_eus` node. Once the answer is received, the text is matched with our dictionary.

On the other hand, the robot control system must be defined, i.e. the meaning of the voice orders must be translated to actions. *MariSorgin* is a synchro-drive robot and as such, two degrees of freedom can be controlled: linear velocity and angular velocity. Thus, the orders that can be given are limited to moving forward/backward, rotating left/right, stopping and accelerating/decelerating. Figure 11 shows how the system is distributed and communicated over the net.

Although in a first attempt linear and angular velocities could be set independently, that is, setting the linear velocity wouldn't affect the current angular velocity (and vice versa), we found that controlling the robot in that manner was rather complicated and that a high level of expertise was needed. Thus, in the final prototype linear and angular velocities are not independently assigned. Modifications of the angular velocity imply that linear velocity is set to zero, and vice versa.

A Qt interface has been developed using `qt` that shows the state of the speech recognition process and the velocity values at each time step. The interface includes minimum distances to obstacles at front, left and right sides, obtained from the laser readings, and the image captured by the robot so that the operator can see what the robot is facing to. Figure 12 shows what this simple interface looks like.



**Fig. 12** *MariSorgin* teleoperation window



**Fig. 13** Experimental setup

*System evaluation:* In order to measure the suitability of the system an experiment has been designed and performed in which 5 persons (3 males and 2 females), all but one not directly involved in the development of the system, were told to give the robot the oral instructions necessary to make the robot reach a predetermined goal from a starting position (see Fig. 13), and results can be seen in Table 2. The theoretical minimum number of instructions refers to the number of steps required by the designed trajectory (forward, left, forward, right, forward and stop). Besides, the empirical min number of instructions refers to the real minimum number of steps done by one of the volunteers.



**Table 2** Results

Theoretical min number of instructions needed	6
Empirical min number of instructions	6
Mean num. of instructions per trip	9.2
Percentage of correctly understood instructions	79 %
Mean time needed to reach the goal	2 min 30s
Minimum time required to reach the goal	1 min 38s

The results of the experiment are not quite significant. The only meaningful thing that can be said is that after a period of training the robot can be operated properly in a real environment. But *MariSorgin*'s laser location is not adequate for obstacle avoidance. The robot will require hard structural changes to get the laser located in an optimal position. This problem is reflected in the teleoperation system, because the obstacle information that the operator can reach does not provide information about table and chair legs, for instance. This could be overcome setting the laser on the old pan-tilt unit and using the *laser\_assembler* package to reconstruct the obstacles laying on the floor and offering the teleoperator the resulting pointcloud. But the main drawback is the delay between the speech identification and the robot action (about 2 s) that makes the system a bit dangerous specially when the robot is speeded up too much, or when the operator does not anticipate enough the order.

Note that it is straightforward to use this package in any of the wheeled robots.

## 6.2 *The nao\_teleop\_speech\_eus Package*

*MariSorgin* is rather limited in its body expressiveness. It is not very appropriate for body language communication. NAO's morphology is much more suitable for HRI and has a huge potential for body language communication and, thus, for exploiting dialogues with humans.

Within the available ROS packages for NAO, the *nao\_speech* node<sup>3</sup> provides the necessary tools for making NAO understand and speak in English. But this node is of no use when another language is required, as it is the case.

Again, the robot control system must be defined, i.e. the meaning of the voice orders must be translated to actions.

A new package named *nao\_teleop\_speech\_eus*) has been developed. Within this package, the *nao\_teleop\_speech* node allows the user to control NAO's movements using several voice commands. The operator, situated in the teleoperation cab (the place where the remote PC is located), gives orders to the robot using a microphone. The robot is able to perform these movements: *Stand up*, *Sit down*, *Move forward*, *Move backward*, *Move left*, *Move right*, *Turn left*, *Turn right* and *Stop*.

---

<sup>3</sup>Developed by M. Sarabia, at the Imperial College London, 2012–2013.

As we previously said, commands are given in Euskara. With the intention to communicate with the robot in a more natural way, the user has more than one choice for each possible command. That is, if the operator wants the robot to stand up, he can say: “Altxatu”, “Tente jarri”, “Zutik jarri”, etc. Therefore a dictionary of some predefined words has been created, including several synonymous for each command. When the user gives a voice command (it can be a long sentence), the voice is converted to text, and processed afterwards. The system tries to find matches between the dictionary and the text received from Google Speech Service. If a match is found, the robot performs the movement corresponding to the received command, otherwise the robot says that it could not understand the order and asks the user to repeat it.

Thus, `nao_teleop_speech` is in charge of receiving messages from `gspeech_eus`, finding any matches in the predefined commands dictionary and deciding which is the action that NAO must perform. It has a subscriber to receive messages that `gspeech_eus` publishes on the `google_speech` topic, and two publishers; one to set NAO’s walking velocity according to the given command, and another one to publish the text messages that NAO has to say.

In order to test the speech capabilities in a HRI context, we integrated the ASR and TTS modules in our Bertsobot project [1]. The Bertsobot project was showed to the general public in a live performance entitled “Minstrel robot: science or fiction”<sup>4</sup> in wich NAO robot showed his verse-improvisation and speech-based communication capabilities. *ZientziaClub* or Club of Sciences is an initiative that aims to disclose science and technologies to the society. RSAIT showed some advances in human-robot interaction by presenting a monologue with NAO: <https://www.youtube.com/watch?v=NEiDw\discretionary-JBER9M>.

We are now working on improving NAO’s body language while speaking in order to show a more human-like behavior and to be more emphatical. For the same reason, NAO’s verbal communication capabilities should be improved so that it could give the same semantical answer using sentences of different gramatical structures, and of course, to perceive feedback from the public or the interlocutor and express accordingly.

## 7 Conclusions

In this paper, some ROS packages have been described and some of the applications given to those nodes were more deeply explained as case studies in concrete robot platforms. Of course, all the developed applications are setup for the rest of the robots. Some videos of life shows can be seen in RSAIT’s youtube video channel (<https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ>) and in our website.

---

<sup>4</sup><http://www.badubada.com/badubadatzen/es/robot-bertsolaria-zientzia-ala-fikzioa/>.

ROS has provided us a tool to standardize this society of robots, different in nature and with different hardware, and has given us the opportunity to set up the same programming and control environment for all the robots. The decision to setup all our robots with ROS allows us to more easily understand, use and maintain them.

It has been hard to reach the actual state. It took time to setup all the robots, to develop the missing drivers and to establish a uniform configuration for all of them. ROS versioning has been a drawback. But it has been worth. Now, it is rather easy to adapt a behavior/application to a different robot. New lab members/students adapt rather quick to ROS basics and can work with any of the platforms. No need to learn several APIs and software environments, neither to know hardware differences among the robots further than movement restrictions and sensor nature.

Thus, rather than a programming tool, ROS has become a methodology for research in robotics. We are willing for ROS 2.0 to have a network of robots communicating to each other and performing operational work inside the faculty.

**Acknowledgments** This work was supported by the Basque Government Research Team Grant (IT313-10), SAIOTEK Project SA- 2013/00334 and the University of the Basque Country UPV/EHU (Grant UFI11/45 (BAILab)).

## References

1. A. Astigarraga, M. Agirrezabal, E. Lazkano, E. Jauregi, B. Sierra, Bertobot: the first minstrel robot, in *Human System Interaction*, (2013), pp. 129–136
2. A. Astigarraga, E. Lazkano, B. Sierra, I. Rañó, I. Zarauz, Sorgin: A Software Framework for Behavior Control Implementation, in *14th International Conference on Control Systems and Computer Science (CSCS14)*, (Editura Politehnica Press, 2003)
3. J. Badger, M. Diftler, S. Hart, C. Joyce, Advancing Robotic Control for Space Exploration using Robonaut 2, in *International Space Station Research and Development*, (2012)
4. G. Ceccarelli, A. Patriti, A. Bartoli, A. Spaziani, L. Casciola, Technology in the operating room: the robot, *Minimally Invasive Surgery of the Liver* (Springer, Milan, 2013), pp. 43–48
5. C. Doarn, K. Hufford, T. Low, J. Rosen, B. Hannaford, Telesurgery and robotics. *Telemed. e-Health* **13**(4), 369–380 (2007)
6. G. Du, P. Zhang, J. Mai, Z. Li, Markerless kinect-based hand tracking for robot teleoperation. *Int. J. Adv. Robot. Syst.* (2012)
7. M.E. Foster, M. Giuliani, A. Isard, C. Matheson, J. Oberlander, A. Knoll, Evaluating Description and Reference Strategies in a Cooperative Human-robot Dialogue System, in *IJCAI*, (2009), pp. 1818–1823
8. Ingenia motion control solutions (2008). <http://www.ingeniamc.com/En>
9. E. Jauregi, I. Irigoien, B. Sierra, E. Lazkano, C. Arenas, Loop-closing: a typicality approach. *Robot. Auton. Syst.* **59**(3–4), 218–227 (2011)
10. J. Koenemann, M. Bennewitz, Whole-body Imitation of Human Motions with a NAO Humanoid, in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (IEEE, 2012), pp. 425–425
11. I. Leturia, A.D. Pozo, K. Arrieta, U. Iturraspe, K. Sarasola, A. Ilarraza, E. Navas, I. Odriozola, Development and Evaluation of Anhitz, a Prototype of a Basque-Speaking Virtual 3D Expert on Science and Technology, in *Computer Science and Information Technology, 2009. IMCSIT'09*, (2009), pp. 235–242
12. H. Mallot, M.A. Franz, Biomimetic robot navigation. *Robot. Auton. Syst.* **30**, 133–153 (2000)

13. M.J. Matarić, *The Robotics Primer* (MIT Press, 2009)
14. D. Matsui, T. Minato, K. MacDorman, H. Ishiguro, Generating Natural Motion in an Android by Mapping Human Motion, in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005. (IROS 2005)*, (IEEE, 2005), pp. 3301–3308
15. F. Mohammad, K. Sudini, V. Puligilla, P. Kapula, Tele-operation of robot using gestures, in *7th Modelling Symposium (AMS)*, (2013)
16. K. Nagatani, S. Kiribayashi, Y. Okada, K. Otake, K. Yoshida, S. Tadokoro, T. Nishimura, T. Yoshida, E. Koyanagi, M. Fukushima, S. Kawatsuma, Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots. *J. Field Robot.* **30**(1), 44–63 (2013)
17. I. Rañó, Investigación de una arquitectura basada en el comportamiento para robots autonomos en entornos semiestructurados, Ph.D. thesis, University of Basque Country, UPV/EHU, (2003)
18. I. Rodriguez, A. Astigarraga, E. Jauregi, T. Ruiz, E. Lazkano, Humanizing NAO robot teleoperation using ROS, in *Humanoids*, (2014), pp. 179–186
19. RWI (1995). Beesoft user's guide and reference. <http://mobilerobotics.cs.washington.edu/docs/BeeSoft-manual-1.2-2/beemanx.htm>
20. SAMPA, Speech assessment methods phonetic alphabet. EEC ESPRIT Information technology research and development program, (1986)
21. A. Setapen, M. Quinlan, P. Stone, Beyond Teleoperation: Exploiting Human Motor Skills with Marionet, in *AAMAS 2010 Workshop on Agents Learning Interactively from Human Teachers (ALIHT)*, (2010)
22. H. Song, D. Kim, M. Park, J. Park, Teleoperation Between Human and Robot Arm using Wearable Electronic Device, in *Proceedings of the 17th IFAC World Congress* (Seoul, Korea, 2008), pp. 2430–2435
23. W. Song, X. Guo, F. Jiang, S. Yang, G. Jiang, Y. Shi, Teleoperation Humanoid Robot Control System Based on Kinect Sensor, in *2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2 (IEEE, 2012), pp. 264–267
24. H.B. Suay, S. Chernova, Humanoid Robot Control using Depth Camera, in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (IEEE, 2011), pp. 401–401
25. R.Y. Tara, P.I. Santosa, T.B. Adji, Sign language recognition in robot teleoperation using centroid distance Fourier descriptors. *Int. J. Comput. Appl.* **48**(2), 8–12 (2012)
26. S. Thrun, W. Burgard, D. Fox, *Probabilistic Robotics*, (MIT Press, 2005)
27. R.T. Vaughan, Massively multi-robot simulations in stage. *Swarm Intell.* **2**(2–4), 189–208 (2008)
28. B. Wang, Z. Li, N. Ding, Speech Control of a Teleoperated Mobile Humanoid Robot, in *IEEE International Conference on Automation and Logistics*, (IEEE, 2011) pp. 339–344

## 11.2 GidaBot: a system of heterogeneous robots collaborating as guides in multi-floor environments

**Title:** GidaBot: a system of heterogeneous robots collaborating as guides in multi-floor environments

**Authors:** O. Parra, I. Rodriguez, E. Lazkano, T. Ruiz


**Journal:** International Journal of Advanced Robotic Systems (IJARS)

**Publisher:** SAGE journals

**Status:** Submitted

**Year:** 2017

# GidaBot: a system of heterogeneous robots collaborating as guides in multi-floor environments

Journal Title  
XX(X):1-11  
© The Author(s) 2017  
Reprints and permission:  
sagepub.co.uk/journalsPermissions.nav  
DOI: 10.1177/ToBeAssigned  
www.sagepub.com/  


## Abstract

GidaBot is an application to setup and run a heterogeneous team of robots for acting as tour guides in multi-floor buildings. Although the tours can reach several floors, robots are not allowed to use the lift and thus, each guiding task requires collaboration among several robots, one per floor. The developed system is based on ROS and makes use of a robust inter-robot communication to share goals and paths during the guiding tasks. A user friendly GUI helps untrained users or new visitors to easily choose target locations or define a list of locations to be visited sequentially. The system robustness has been tested using real robots at the Faculty of Informatics in San Sebastian. The whole application is available together with a simulated world so that the system functioning can be checked further.

## Keywords

Service robots, Tour guide robots, Distributed robotic system

## Introduction

Robots are moving from industries to other locations to be part of our daily lives, helping us performing different tasks. With the advance of robotics technologies, researchers have started to explore the applications of service and social robots. Service robots – defined as robots that perform useful tasks for humans [International Federation of Robotics \(accessed January 24, 2017\)](#)– have many and diverse applications, such as assisting humans in transportation (self-driving cars), vacuum cleaners used in home environments [Forlizzi and DiSalvo \(2006\)](#) (iRobot's Roomba) or for commercial use (Sealed Air's Intellibots [Sealed Air – Diversey Care \(accessed January 26, 2017\)](#)), drones for photographing and transportation, nursing robots (Pearl [Pineau et al. \(2003\)](#) and RIBA – the friendly robot nurse [Riken-Tri \(accessed January 24, 2017\)](#) [Mukai et al. \(2010\)](#)), care assistants for the elderly at home [Fischinger et al. \(2016\)](#) or robots acting as shopping assistants [Kanda et al. \(2009\)](#).

A different scope of application that also relies on heavy autonomous navigation capabilities is that of tour-guide robots. The research presented here focuses on a heterogeneous robot navigation system that enables robot communication for cooperative guiding tasks in different floors, and allows individual navigation in each floor at the same time. In this particular case, we use four mobile robots available in our robotics research group. Albeit the system has been set up to solve the multi-floor navigation problem of the Faculty of Informatics in San Sebastian, the system can be adapted to a different building and robot configuration.

The paper is structured as follows. Section reviews the literature and emphasizes the advantages of a distributed robot collection for tour guiding in multi-floor environments. Next, Section summarizes the basic navigation capability that a robot acting as a guide needs and the GUI designed for interacting with the user. This system is extended in Section in order to develop a distributed heterogeneous

guide-robot system. Features such as the system design, the expansion of the GUI and the information exchange among the involved robots are described, together with the system setup in a ROS based software architecture.

The designed system has been tested in a robot/environment configuration described in Section and results are shown in Section. The paper ends as corresponds with Section dedicated to conclusions and the description of the future work.

## Tour guide robots

The literature review reveals several instances of robots acting as tour guides. Minerva [Thrun et al. \(1999\)](#) is very likely the first robot that acted as such in the Smithsonian's National Museum of American History in Washington, and by far the most cited one. In [Rosenthal et al. \(2010\)](#) the navigation capabilities of CoBot are evaluated while acting as a guide through a cooperation between the visitor and the robot, helping each other to fulfill the task. More recently, a robot that performs guided tours was designed, built and set up at the Eureka Science Museum of San Sebastian [Susperregi et al. \(2012\)](#). Some authors emphasize the need for social interaction in such platforms. Robovie assisted visitors at the Osaka Science Museum Exhibit [Shiomi et al. \(2007\)](#), and the humanoid robot Robotinho [Faber et al. \(2009\)](#), mounted on a wheeled platform to reduce mobility constraints, showed such capabilities while acting as a guide in the Deustches Museum of Bonn.

[Trahanias et al. \(2010\)](#) present a different approach in which robots are teleoperated over the internet and act as interactive agents in populated environments as museums and exhibitions. In addition, [Hristoskova et al. \(2012\)](#) propose a distributed collaboration between two robots acting as guides. Robots share profiles and tour information with the aim of automatically exchanging the group members in order



to optimize the amount of interesting content each time robots are in the neighborhood.

Looking forward in a near future, one crucial challenge to be considered is the usefulness of those robots –initially intended to be used in single floor buildings– in public places with multiple floors. A possible solution to the multi-floor navigation problem is the use of a single robot that can navigate through different floors using lifts, as the robot Charlie Troniak et al. (2013) is able to do. A similar work extended to multiple robots, also using elevators, is proposed in the GuideBot tour guide López et al. (2013a) and BellBot hotel assistant López et al. (2013b) systems. But entering lifts may be dangerous for robots, depending on the security measures, the gap on the floor, the robot’s geometry, and specially, the drive system. Also, robots that get into lifts are supposed to have the necessary abilities to interact with the lift interface, inside and outside, to execute precise actions. The lack of proper actuators can be overcome by interacting with humans as CoBot does Rosenthal et al. (2010) Veloso et al. (2012). This symbiotic collaboration approach has been further expanded to a homogeneous team of up to 4 robots that are also able to perform delivery tasks Veloso et al. (2015). An alternative is to use multiple robots, limiting the work scope of each robot to a single floor and avoiding the robots using the elevators. Although several platforms are needed, robot navigation is more secure (robot paths don’t collide) and several tours can be run concurrently. In this way, the lift remains available for people involved or not in the guided tour and for people with reduced mobility. Of course, robust robot communication must be guaranteed in order to achieve a successful system.

### Single robot guide system

This section briefly describes the features of the ROS based tour-guide robot system developed for a single robot and in which the multirobot guide system is based on.

#### Robot navigation setup

A robust guide system mainly relies on robust navigation capabilities. For any service robot it is fundamental to be able to safely and accurately navigate in its environment, and so it is for guide robots. Robot navigation implies that the robot is able to determine its own position and plan a path towards some goal location, avoiding dangerous situations, such as collisions with surrounding objects in the environment. AI techniques make use of probability distributions to represent and maintain uncertainty in robot localization and feature identification over time in order to determine the path the robot must follow to fulfill a task Thrun et al. (2005). These probabilistic approaches have shown to perform well in semi-static environments, and thus, are being widely used in multiple robotic systems. ROS\* provides a navigation stack initially developed for the PR2 robot by Willow Garage Marder-Eppstein et al. (2010) that has been adapted for many robots<sup>†</sup>. This navigation stack offers tools for constructing a global map of the robot’s environment by means of SLAM (Self Localization and Mapping) techniques. Besides, robot localization during navigation is maintained using particle filter based AMCL (Augmented Monte Carlo Localization) algorithm. Together

with the map and a robot localization mechanism, the navigation stack needs a planner to find and select the path to be followed by the robot. ROS allows the user to configure the stack by choosing among several planners the one that better fits to the robot/environment system. The default navigation function makes use of Dijkstra’s algorithm for planning purposes.

#### Robotic platform

*Kbot* is a differential drive robot supplied with a Sick S3000 laser scanner built by Neobotix Neobotix ([accessed January 24, 2017]) in 2004 for acting as a tour guide at the Eureka Museum of Science in San Sebastian. It has been recently renewed, with a new PC, a Kinect sensor and a smaller touch screen mounted on it. This robot is now ROS operative, since the necessary ROS drivers have been developed for it Rodriguez et al. (2016).

#### Graphical User Interface

Tour-guiding robots need to interact with humans. The simplest way of interaction requires a graphical user interface (GUI) so that tasks and goals can be input by the user and information can be feedbacked to him/her in a practical though not human-like manner. Such interface has been developed using Qt. Fig. 1 shows how the initial GUI looked like. The most frequently demanded sites of our faculty are located at the first floor. Hence, in this attempt a map of the first floor has been created with ROS mapping utilities and that map is being used as the floor plan of the developed GUI. This floor plan has been populated with several interaction buttons corresponding to the important locations people might be interested in, such as the secretary’s office, the Dean’s office, the lift, several labs and so on. The robot morphology makes door crossing insecure and thus, for the time being the GUI limits the robot guiding task to the front of the door that gives access to the desired location. Information about current and destination locations is also displayed on the interface. Specifying the robot’s initial pose for the ROS navigation stack is RVIZ<sup>‡</sup> dependent, so the system operator is needed to perform this task.

This interface together with the underlying navigation system were tested in *Kbot* (Fig. 2) with different visitors and bachelor student groups that came to the faculty for the first time Rodriguez et al. (2016).

#### Gidabot: multirobot guide system

The single robot navigation setup explained before has been used as a base for developing the interfloor multirobot guide system described herein. The developed guide robot was limited to a single floor, so the next goal was to extend the guide system to be able to allow tours all along the different floors of the building.

In order to satisfy users’ requirements (move to a goal or follow a goal sequence) the system has to provide different

\*<http://www.ros.org>

<sup>†</sup><http://wiki.ros.org/navigation/RobotsUsingNavStack>

<sup>‡</sup><http://wiki.ros.org/rviz>



Figure 1. Kbot's navigation interaction window and the current local costmap in RVIZ



Figure 2. Snapshot of a guided tour with bachelor student visitors

operation modes. Moreover, the content of the information shared among robots, and the way it is transmitted have to be correctly defined, ensuring a robust and efficient robot communication.

### System design

The multirobot system has been designed to allow two operation modes:

**Single target mode** In single target mode the user can only select a single target from all available locations. The robots must cope with different situations:

1. The user and the desired goal are on the same floor. In this situation, only one robot will guide the user from the beginning to the end of the navigation. Therefore, the only action the robot must perform is to reach the goal.
2. The user and the desired goal are on different floors. In this case two robots are involved in guiding tasks; one is located in the floor where the trip starts and the other one in the floor where the navigation ends.
  - (a) In the floor where the navigation starts, the robot will guide the user from the starting location to the lift or staircase (the user is free to choose) and it will indicate what floor to go. Another robot

will be waiting for the user in the goal floor's meeting point to solve the remaining path to the goal location.

- (b) In the floor where the navigation ends, the robot needs to meet the user before guiding him/her, so it has to move to the previously chosen lift or staircase (the meeting point). Afterwards, when the user arrives, he/she will notify the robot to go on. Then, the robot will guide him/her to the goal and the navigation will finish.
3. The robot is not involved in the navigation task. If the robot is not in the initial floor neither in the goal floor, it does not have to do anything; just wait to receive the next goal where it is involved.

If a robot receives more than one request, these are queued in order of arrival and managed using a First-In-First-Out (FIFO) queue. Thus, pending goals (with initial or final point in the robot's floor) are processed in the same order they are requested. Once a robot has finished processing a request, if its queue is not empty, it will start navigating to the next goal (the first in the queue).

### Tour mode

Often, it can be interesting to follow a predefined goal sequence, for instance, to show the surroundings. This means that robots will conduct guided tours. For this purpose, the system allows to create tours as a collection of location goals in the desired sequence. Tours are saved in a local directory and can be edited. New and edited tours are automatically shared among all the available robots involved in the tour.

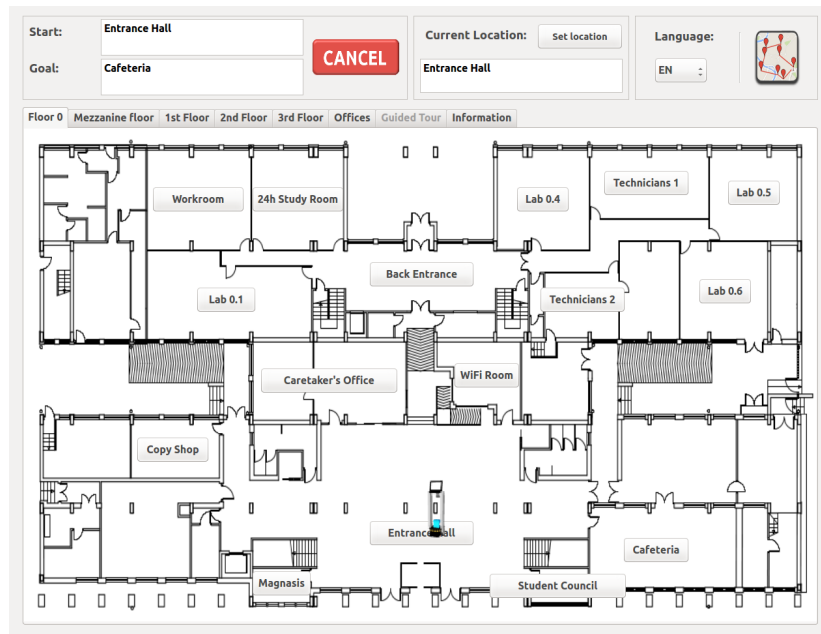
Of course, the *Tour mode* relies on the *Single Target mode* and can face the same situations at each target of the sequence.

### Expanding the GUI

The previously developed GUI was designed for a single robot navigation, and limited for a single floor. Extending the system to allow guided tours among different floors requires improvements in usability and intuitiveness of the GUI to satisfy the users' needs. Fig. 3 shows the renewed aspect of the interface.

The graphical interface is available in Basque, English and Spanish languages, and offers the user the option to select





**Figure 3.** Main view of the renewed GUI

between the single target and tour modes by clicking the upper right button with a path with waypoints.

The main window comprises a tab per floor, showing buttons of the most interesting locations and the current position of the robot on the floor's blueprint, which replaces the original map created with ROS tools. Three more tabs (*Offices*, *Guided Tour* and *Information*) together with the current navigation information complete the interface. The *Offices* tab shows information about the available destination points per floor, in order to help the user find the desired target locations. The *Guided Tour* tab allows the user to select and follow a predefined goal sequence. And the *Information* tab offers information about operative robots and detailed info about velocities and battery level of the current robot.

In single target operation mode, after selecting the tab corresponding to the floor of interest, the user has to click on the destination button, accept the confirmation message and choose the way to move between floors (if there is a floor change). Then, the robot will start moving and the GUI will show its location on the floor's map during the whole trip. Each time the user wants to send a goal, the GUI informs him/her about the number of pending requests of the goal robot. Therefore, if he/she thinks it will take long to wait, can choose to cancel the task. Moreover, the *Information* tab also shows the number of pending requests of each operative robot.

The *Guided Tour* tab (Fig. 4) offers the user the option to select and follow an already predefined tour. In tour mode, the user can start a tour by clicking *Start again* button, and when he/she reaches a goal and is ready to go on, the *Continue* button must be clicked on. In case the user wants to skip any goal, the next desired one must be selected from the tour site list.

In both operation modes, if the user wants to finish the navigation on the way to the goal, the "Cancel" button must be clicked. The robot will stop and it will immediately become ready to process a new goal. In case the navigation comprises several floors, the rest of the affected robots' navigation will also finish. Besides, if a goal is unreachable, for instance, when every possible path is closed, the robots involved in the navigation will be informed and the user will be told that the robot can not help him/her.

Besides, the system always keeps the user informed about the trip and the actions needed to perform, via text pop-ups and verbal messages. For instance, the robot notifies the user when the target point is reached, or in case of a floor change, it informs about the floor he/she has to move to in order to continue with the trip.

To address the inconvenience of setting manually the initial pose of the robot with RVIZ, now the graphical interface provides the option to indicate the robot's position and orientation in a simple manner, using the current floor's map in the GUI (see Fig. 5). This involves calculating and modifying the current location and updating the information text boxes, considering the map's scale and consequent precision loss.

### *Information exchange*

The multirobot guide system proposed here is mainly designed to work with several robots interconnected carrying out collaborative navigation tasks, but single navigation is also supported. When multiple robots are involved information exchange among the robots is required.

No matter how many choices the system offers, if the system is to be robust and efficient, then it is mandatory to ensure a proper and reliable communication among robots. In order to make the system work as desired, each robot has to inform the others, on the one hand, about user's requests

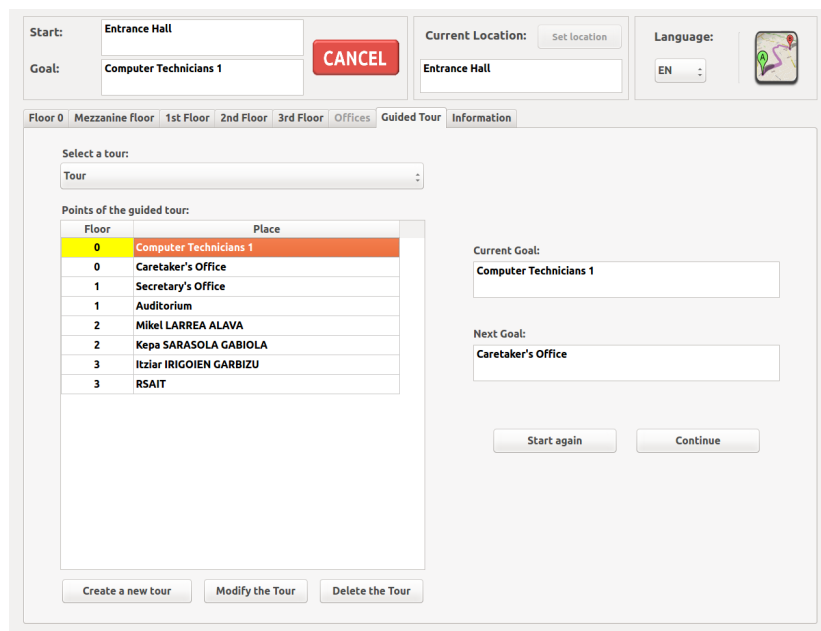


Figure 4. GUI tab for guided tours

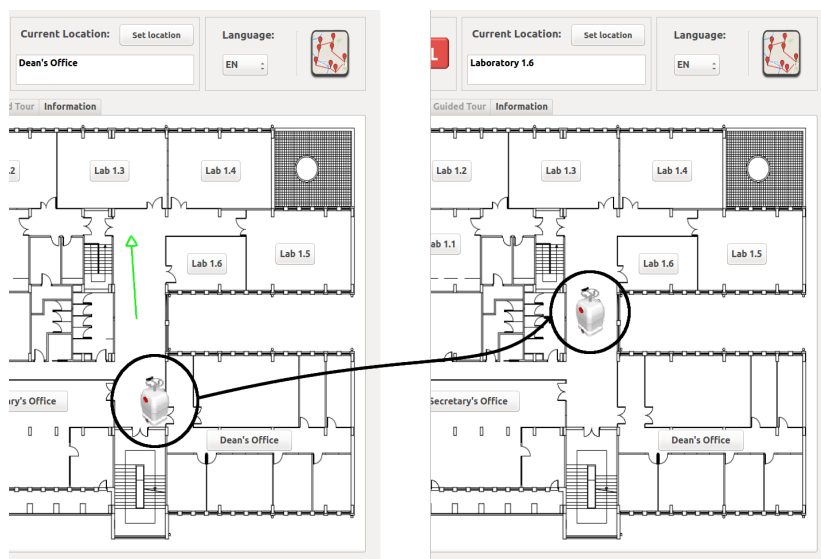


Figure 5. Setting up the initial pose. The green arrow on the left is drawn with the mouse after clicking the “Set Location” button. The figure’s right side shows how the interface updates the robot’s position

and, on the other hand, about its state – current location and navigation state. This communication must be fluent over time.

Context exchange among robots relies on different type of messages explained below.

**Goal descriptions** A goal comprises the information related to the start and end points of the navigation. Note that one or several robots can be involved in guiding tasks:

- Number of the initial floor, where the visit starts.
- Coordinates of the initial robot location.
- Number of the goal floor, where the visit ends.
- Coordinates of the goal location.

- Way to be taken, chosen by the user. When the initial and goal locations are in different floors the user must select either the lift or the staircases, so that the robots responsible for the navigation leave and meet the user at the correct location.
- Start point identifier.
- Goal point identifier.
- Language to be used for (verbal and text) communication.

**Tour description** Tours comprise a predefined goal sequence and are defined as local text files in a two column format with the following content: *floor; room identification*. Optionally,

a comment can be added preceded by a # character. Here an example of a tour definition text file:

```
Tour
0, 014 #Computer Technicians 1
0, 008 #Caretaker's Office
1, 101 #Secretary's Office
1, 104 #Auditorium
2, 223 #Mikel LARREA ALAVA
2, 274 #Kepa SARASOLA GABIOLA
3, 325 #Itziar IRIGOIEN GARBIZU
3, 305 #RSAIT
```

When several robots are involved in a new tour, the tour is shared among the platforms involved, which mean each robot has to update the list of available tours and show it in the corresponding tab of its GUI. The tour information is shared via a message containing the following fields:

- Robot id: robot where the tour was created.
- Tour name: the name given to the tour.
- Tour file name: the name of the file where the tour is saved.
- Tour information: contains the goal sequence information, in string format.

**Robots' pose** As the GUI shows every available robot's localization in the map, it needs to update each robot's current position in the corresponding tab. Thus, robots continuously interchange their location provided by the AMCL algorithm:

- X: the X coordinate of the robot.
- Y: the Y coordinate of the robot.
- Orientation: the robot orientation in quaternion form.

**Pending requests** As mentioned before, when a user needs to be guided to a destination, she/he is informed about the pending requests of the other robot involved in the task, so that she/he can decide to abort the task or to go on. This type of messages are defined as:

- The number of pending requests.
- The list of pending goals.

### ROS setup

The main part of the interfloor multirobot guide system is based on a single ROS node named *multirobot\_navigation*, which receives navigation goals, processes this information and then sends the robot to the pertinent place. The *multirobot\_navigation* node is executed in each robot, and depending on the information of the received goal messages, it responds in a way or another, as explained in section .

Messages (*multirobot goals*, *tours*, *navigation poses* and *pending requests*) are broadcast to all the available robots and each one decides whether the received information is relevant or it can be ignored. ROS is not designed for multiple robot systems where information must be shared among all the entities, although there are two packages that facilitate a solution. Our first attempt was to use the *multimaster\_fkcie* package, which allows to establish and manage a multimaster network using multicast protocol,

but our faculty's network firewall does not allow many-to-many distribution. For that reason, our system was developed using the *multimaster* package, which – though deprecated – enables communication between two ROS masters. What it does, exactly, is to register topics or offer a service at a different ROS master, and/or subscribe to topics or call to services of the same master. In this manner, topics/services managed by different masters are selected to be shared.

In the developed system, each robot has its own ROS master and all robots are interconnected on a complete graph network, resulting in a low latency messaging platform. In order to share information using the *multimaster* node, it is enough to execute this node in just one of the two masters we want to connect. This means we need  $\frac{n(n-1)}{2}$  multimasters, where  $n$  is the number of robots. Before running the node, the foreign master must be specified, together with the local publications we wish to share and the foreign publications to be received.

Fig. 6 shows the system architecture for 4 robots. Its configuration will be described in Section . Notice the multimasters on the bottom row.  $M_{ij}$  denotes that the multimaster communicates masters  $i$  and  $j$ , i.e. robots  $i$  and  $j$ . For the experimental setup used in this work, 6 multimasters are needed in order to have full intercommunication among the 4 robots.

Table 1 shows the context exchange in terms of ROS messages.

**Table 1.** Messages description

Message Type	Data type	Name
Goal	float32	initial_floor
	geometry_msgs/Point	initial_pose (x, y)
	float32	goal_floor
	geometry_msgs/Pose	goal_pose (x, y, $\theta$ )
	string	way
	string	start_id
	string	goal_id
Tour	string	language
	float32	floor
	string	tour_name
	string	file_name
Pose	string[]	tour_goals_info
	geometry_msgs/Pose	nav_pose (x, y, $\theta$ )
Pending requests	uint8	num
	Goal[]	goals

### Experimental setup

The developed guide system has been tested with four different robots moving through the floors of the Faculty of Informatics. The experimental setup is described below.

#### Robotic platforms and environment description

As mentioned in Section , *Kbot*, a differential drive robot supplied with a Sick S3000 laser scanner, was the first platform to setup and the base for adapting and configuring the other available mobile platforms.

*MariSorgin* is our heirloom robot, a synchro-drive robot that dates from 1996. It is a B21 model from Real World

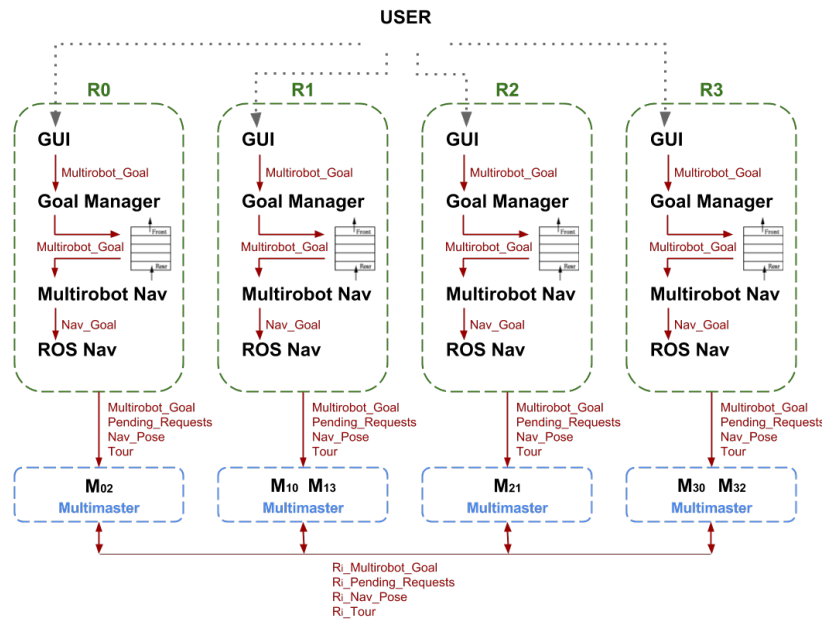


Figure 6. Multirobot system architecture

Interface provided with a ring of ultrasound, infrared and tactile sensors for obstacle avoidance. In addition, a Hokuyo URG-30 laser, a Kinect sensor and a Heimann thermopile have been placed on top of the enclosure.

*Galtxagorri* and *Tartalo* are two differential drive robots, a Pioneer-3DX and PeopleBot models from Omron Adept MobileRobots **Omron Adept MobileRobots** (accessed January 24, 2017) respectively. Both are provided with ultrasounds and a Cannon VCC5 camera, the former has a Leuze RS4 mounted on its top and the later a Sick LMS200 laser sensor on its base.

Summarizing, all the robots have a laser sensor for safe navigation and localization, a touch screen for accepting user requests and speakers to be able to reproduce audio. This brief description gives a hint of the diversity of sensors being used and the dissimilar morphology and sizes, in summary the heterogeneity of the robot team.

Regarding the environment, the Faculty of Informatics of the University of the Basque Country (UPV/EHU) is located in San Sebastian. It is a five floor building equipped with two side staircases and a single lift that enable people to move between these floors.

The main entrance is in the zeroth and lowest floor, where a few lecture rooms and laboratories are placed, just like in the mezzanine. The mezzanine floor is peculiar in the sense that it has no public lift access. The Dean's Office, Secretary's Office and Auditorium can be found in the first floor, together with more lecture rooms and laboratories, whereas most professors' offices are located in the second and third floors. Research laboratories are also in this upper floor.

### Wireless communication

In order to be able to communicate over time, either robots are in the same LAN or they have a known public IP address

externally accessible. The Faculty of Informatics is a public building of the University and thus, wireless communication options are preset and not all possibilities are authorized. The following alternatives have been considered:

- LAN: requires multiple antennas. Not available/authorized.
- *eduroam* (education roaming)<sup>§</sup>: although there are several antennas distributed all over the building, in its current state the connection suffers multiple interruptions and is not reliable at all.
- 3G Mobile Wi-Fi with prepaid SIM cards: occasional communication interrupts may occur, but they are rare. Hence, this is the final choice we made. Each robot now uses a 3G Modem that connects to the Internet.

With this setup, considering that the robots share information among each other, they need to have an accessible IP address assigned so that messages can be received properly. This IP address should not be changed while the system is operating, and preferably, neither after each session.

However, the used SIM cards do not allow static IPs; IPs can change at any time, and even more, those IPs are not externally accessible. So after discarding the choice of hiring a static IP Internet connection with a telephone provider, we decided to set up a Virtual Private Network (VPN) that enables assigning static IPs to the robots.

As the final solution, we managed to get four static IP addresses from our university's VPN service, creating an LDAP<sup>¶</sup> account for each robot and using it when connecting to VPN. This way, an accessible static IP address, in which

<sup>§</sup><http://www.eduroam.org>

<sup>¶</sup>Lightweight Directory Access Protocol

the ROS master will be running, is set every time the robots initialize.

### Configuration and setup

The developed system is easily adaptable to different robot placements. Each robot has its own launch file, containing several parameters that must be set up before executing the system, and the initialization of the required modules (Fig. 6).

Among these parameters, the global configuration of the system must be set up. Each robot needs to know the floor it will be working on, and for the other robots participating in the task, their name, IP address and the floor they will be working on. It is also possible to have a floor without any robot, which means that the user will have to find the destination by himself/herself. In our case, all of our robots are prepared to navigate in any floor of the Faculty of Informatics of San Sebastian, so changing their location can be easily carried out.

Regarding robot communication, the robots where the multimaster nodes will be running can also be configured within the launch files.

## Experimental Results

This section describes the different experiments performed to evaluate the robustness of the *GidaBot*. Experiments have been performed in both, simulated and real robot/environment systems.

### GidaBot in simulation

The goal of the simulated experiment is twofold:

1. Evaluate the soundness of the application without suffering from issues like battery life and Wi-Fi communication.
2. Offer a tool to other researches/developers to test the application.

With that aim, each floor of the Faculty of Informatics and each robot has been modelled using Blender<sup>||</sup> and, afterwards, these models have been integrated in five different Gazebo<sup>\*\*</sup> worlds. As a result a complete simulation environment is available together with the *GidaBot* system. Fig. 7 shows screenshots of each simulated floor/robot pair.

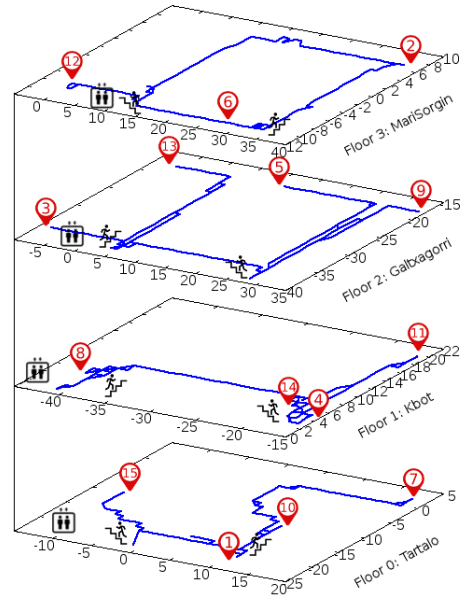
A long tour of 15 randomly selected goals covering the whole faculty was defined and again, the goal sequence was randomly chosen (see Table 4).

Four standard PCs (i5 with 4GB RAM) available at the lab where used. Instead of the individual 3G modems, all the PCs were connected to the same Wi-Fi router that gave internet access. Two people were responding to the interface queries during the experiment that last 1 hour approximately. No errors occurred during the tour.

Fig. 8 shows the paths followed by each robot on each floor during the guided tour. Table 3 summarizes the number of goal messages exchanged among the robots in the course of the tour, only the number of messages received and processed of each robot have been considered. Note that when floor change is required two robots are involved in the task, therefore some goal messages are processed by more than one robot.

**Table 2.** Long duration tour

#	Floor	Place
1	0	Student Council
2	3	IXA Lab
3	2	C. Rodriguez's Office
4	1	Dean's Office
5	2	J. Abascal's Office
6	3	B. Sierra's Office
7	0	Computer Technicians 1
8	1	Ada Lovelace Auditorium
9	2	Seminar Room 2.3
10	0	WiFi Room
11	1	Laboratory 1.3
12	3	RSAIT Lab
13	2	O. Arbelaitz's Office
14	1	Secretary's Office
15	0	Copy Shop



**Figure 8.** Path followed by each robot on each floor during the guided tour

**Table 3.** Exchanged goal messages among the robots during the guided tour

Robot	Received	Processed
<i>Tartalo</i>	15	7
<i>Kbot</i>	15	8
<i>Galtxagorri</i>	15	8
<i>MariSorgin</i>	15	6

A video of the simulated system running a tour is also available at RSAIT's YouTube channel<sup>††</sup>.

<sup>||</sup> <https://www.blender.org/>

<sup>\*\*</sup> <http://gazebo.org/>

<sup>††</sup> <https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ>



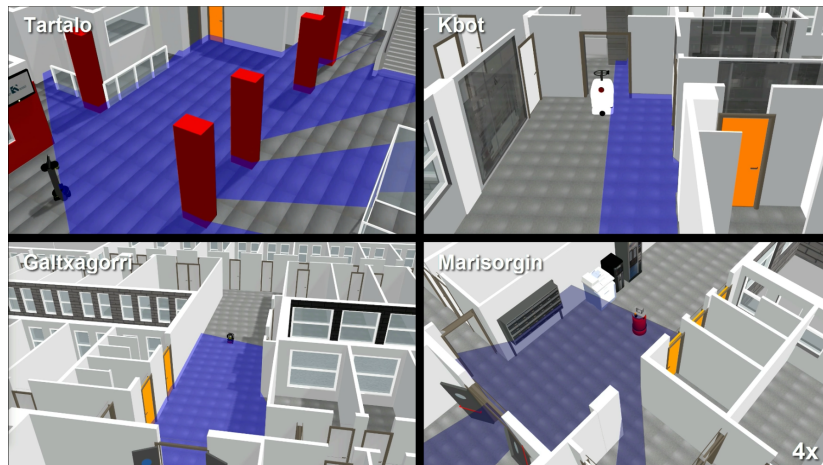


Figure 7. Screenshots of the simulated robot/environment pairs

### GidaBot in the real world

In order to analyze the system performance in the real world, a randomly selected guided tour of eight randomly chosen goals that covers the main four floors of the Faculty of Informatics has been defined:

Table 4. Real world tour

#	Floor	Place
1	0	Technicians' Lab
2	0	Caretakers' Office
3	1	Secretary's Office
4	1	Ada Lovelace Auditorium
5	2	M. Larrea's Office
6	2	K. Sarasola's Office
7	3	I. Irigoien's Office
8	3	RSAIT Lab

The guided tour has been performed by an untrained user on a typical working day. Information about the required time and the traveled distance to successfully complete the tour have been collected during the whole process (Table 5). The mean linear velocity varies depending on the characteristics of the environment, navigational capabilities and configuration parameters of each robot. Note that the time and distance have only been taken into consideration when the user was with the robot.

Table 5. Collected data

Robot	time (s)	dist. (m)	mean vel. (m/s)
Tartalo	353	86	0.24
Kbot	81	32.2	0.4
Galtxagorri	172	60	0.35
MariSorgin	132	35.8	0.27
Tour	738	215	0.29

In addition, a video available at RSAIT's YouTube channel shows the whole tour recorded during the guided tour. The main frame is divided into four subwindows, one per floor and robot. The upper left subwindow corresponds to the zeroth floor and *Tartalo*. The upper right one to the first floor and *Kbot*. The lower left corner to the second floor

and *Galtxagorri*. And the lower right subwindow to the third floor and *MariSorgin*. Fig. 9 shows a snapshot of the video.

The video also shows the pop-up windows displayed to the user at the different steps of the tour, in order to measure the kind of interaction the user has with the robots. As the user chooses to reach upper floors by the stairs, it can be appreciated how the next robot in each phase moves to the stairs when it receives the message from the currently acting robot, in order to meet the user and continue with the tour.

As mentioned before, the configuration of the system can be adapted to a different setup. The code will be available at RSAIT's GitHub<sup>‡‡</sup>.

### Conclusions and further work

The GidaBot system described in this paper is an application to setup and run multiple robots in tour guiding tasks over multi-floor environments. The developed guide system makes use of a robust inter-robot communication among different mobile platforms and allows them to carry out guiding tasks along the different floors of the building. The application also includes a Graphical User Interface that helps the user interact with the robot in an intuitive manner.

It is important to emphasize the heterogeneity of the robot team where the application is being tested. But also the fact that a standard software framework is being used. These two features make possible to conclude that it is not a tour-guide system developed just for a specific robot/environment system, but that it is applicable (with some adjusts) to other instances. This would require some work such as adapting the maps shown on the tabs of the interface and the coordinates of the interesting locations of the new environment.

The system relies on the ROS navigation stack that needs to be tuned on each robot. The performance of the navigation stack varies depending on the hardware. Some parameters could be better tuned in some cases to avoid bizarre behavior such as giving several turns on the spot when localization fails. Also, bothering the robot, blocking

<sup>‡‡</sup><https://github.com/rsait>



**Figure 9.** Snapshot of the real system during a guided tour

its way and preventing robust localization can entail a failure. Further versions of this ROS package may overcome this problem. Meanwhile, users must be warned about it. Visitors also must be advised to keep at the back of the robot to minimize sensor uncertainty and overcome localization problems.

Regarding robot communication, for the application to be run, each robot must have a known static IP and all the robots must share the network. Though the used network resources are irrelevant for the application itself, if the system is going to be used in a public building as it is the case, in a near future, it would be desirable to avoid the use of prepaid SIM cards and make the system run with *eduroam*.

A relevant issue concerns the *multimaster* package used for context sharing among robots. ROS was designed for single robot systems, and the *multimaster* package is a side solution. ROS 2.0 is being designed to overcome this problem offering the possibility of building multirobot systems. But ROS 2.0 version is still in a beta stage and the migration from ROS 1.0 to 2.0 promises to be anything but trivial with many incompatibilities among packages.

As further work, the system can be tuned in several aspects. In larger environments two or more robots could share a floor and then, robot availability should also be managed. But depending on the field of application – and it is the case of tour-guide robots – service robots should be able to interact with users in a human like manner and show social skills. Currently, we are integrating a face recognition system so that single visitors are recognized while sharing a goal among robots. User images must be shared among robots, but first, the degree of acceptability by the potential users must be measured. Besides, we intend to extend the system so that the robots respond to spoken orders.

## References

- Faber F, Bennowitz M, Eppner C, Görög A, Gonsionr C, Joho D, Schreiber M and Behnke S (2009) The humanoid museum tour guide Robotinho. In: *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. pp. 891–896.
- Fischinger D, Einramhof P, Papoutsakis K, Wohlkinger W, Mayer P, Panek P, Hofmann S, Koertner T, Weiss A, Argyros A et al. (2016) Hobbit, a care robot supporting independent living at home: First prototype and lessons learned. *Robotics and Autonomous Systems* 75: 60–78.
- Forlizzi J and DiSalvo C (2006) 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, pp. 258–265.
- Hristoskova A, Agüero CE, Veloso M and De Turck F (2012) Personalized guided tour by multiple robots through semantic profile definition and dynamic redistribution of participants. In: *Proceedings of the 8th International Cognitive Robotics Workshop at AAI-12, Toronto, Canada*.
- International Federation of Robotics ([accessed January 24, 2017]) IFR. <http://www.ifr.org/service-robots/>.
- Kanda T, Shiomi M, Miyashita Z, Ishiguro H and Hagita N (2009) An affective guide robot in a shopping mall. In: *Proceedings of the 4th ACM/IEEE international conference on Human-robot interaction*. ACM, pp. 173–180.
- López J, Pérez D, Santos M and Cacho M (2013a) Guidebot. A tour guide system based on mobile robots. *International Journal of Advanced Robotic Systems* 10.
- López J, Pérez D, Zalama E and Gomez-Garcia-Bermejo J (2013b) Bellbot - a hotel assistant system using mobile robots. *International Journal of Advanced Robotic Systems* 10.
- Marder-Eppstein E, Berger E, Foote T, Gerkey B and Konolige K (2010) The office marathon: Robust navigation in an indoor office environment. In: *International Conference on Robotics and Automation*.
- Mukai T, Hirano S, Nakashima H, Kato Y, Sakaida Y, Guo S and Hosoe S (2010) Development of a nursing-care assistant robot RIBA that can lift a human in its arms. In: *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, pp. 5996–6001.
- Neobotix ([accessed January 24, 2017]) Neobotix - robotics and automation. <http://www.neobotix-robots.com>.
- Omron Adept MobileRobots ([accessed January 24, 2017]) Omron adept Mobile robot systems for research and development. [http://www.mobilerobots.com/Mobile\\_Robots.aspx](http://www.mobilerobots.com/Mobile_Robots.aspx).

- Pineau J, Montemerlo M, Pollack M, Roy N and Thrun S (2003) Towards robotics assistants in nursing homes: Challenges and results. *Robotics and Autonomous Systems* 42: 271–281.
- Riken-Tri ([accessed January 24, 2017]) Riba: World's first robot that can lift up a human in its arms. collaboration center for human-interactive robotic research. <http://rtc.nagoya.riken.jp/RIBA/index-e.html>.
- Rodriguez I, Jauregi E, Astigarraga A, Ruiz T and Lazkano E (2016) *Robot Operating System (ROS). The Complete Reference*, chapter Standardization of a heterogeneous robots society based on ROS. Number 625 in *Studies in Computational Intelligence*. Springer, pp. 289–313.
- Rosenthal S, Biswas J and Veloso M (2010) An effective personal mobile robot agent through symbiotic human-robot interaction. 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, pp. 915–922.
- Sealed Air – Diversy Care ([accessed January 26, 2017]) TASKI' intellibot. <http://www.intellibotrobotics.com/>.
- Shiomi M, Kanda T, Ishiguro H and Hagita N (2007) Interactive humanoid robots for a science museum. *IEEE Intelligent systems* 22(2): 25–32.
- Susperregi L, Fernandez I, Fernandez A, Fernandez S, Murtua I and de Vallejo IL (2012) Interacting with a robot: a guide robot understanding natural language instructions. In: *Ubiquitous Computing and Ambient Intelligence*. Springer, pp. 185–192.
- Thrun S, Bennewitz M, Burgard W, Cremers AB, Dellaert F, Fox D, Hähnel D, Rosenberg C, Roy N, Schulte J et al. (1999) Minerva: A second-generation museum tour-guide robot. In: *Robotics and automation, 1999. Proceedings. 1999 IEEE international conference on*, volume 3. IEEE.
- Thrun S, Burgard W and Fox D (2005) *Probabilistic robotics*. MIT press.
- Trahanias P, Burgard W, Argyros A, Hähnel D, Baltzakis H, Pfaff P and Stachniss C (2010) TOURBOT and WebFAIR: Web-operated mobile robots for tele-presence in populated exhibitions. *IEEE Robotics and Automation Magazine* 12(2): 77–89.
- Troniak D, Sattar J, Gupta A, Little JJ, Chan W, Calisgan E, Croft E and Van der Loos M (2013) Charlie rides the elevator—integrating vision, navigation and manipulation towards multi-floor robot locomotion. In: *Computer and Robot Vision (CRV), 2013 International Conference on*. IEEE, pp. 1–8.
- Veloso M, Biswas J, Coltin B and Rosenthal S (2015) Cobots: Robust symbiotic autonomous mobile service robots. In: *24th International Joint Conference on Artificial Intelligence*. pp. 4423–4428.
- Veloso M, Biswas J, Coltin B, Rosenthal S, Brandão S, Mericli T and Ventura R (2012) Symbiotic autonomous service robots for user-requested tasks in a multi-floor building. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*.



### 11.3 Humanizing NAO robot teleoperation using ROS

- Title:** Humanizing NAO robot teleoperation using ROS
- Authors:** I. Rodriguez, A. Astigarraga, E. Jauregi, T. Ruiz, E. Lazkano
- Conference:** International Conference on Humanoid Robots (Humanoids)
- Publisher:** IEEE
- DOI:** 10.1109/HUMANOIDS.2014.7041357
- Year:** 2014

# Humanizing *NAO* robot teleoperation using ROS

Igor Rodriguez<sup>1</sup>, A. Astigarraga<sup>1</sup>, E. Jauregi<sup>1</sup>, T. Ruiz<sup>1</sup>, E. Lazkano<sup>1</sup>

**Abstract**—The work presented here proposes two different ROS packages to enrich the teleoperation of the robot *NAO*: speech-based teleoperation (in Basque) and gesture-based teleoperation together with arm control. These packages have been used and evaluated in a human mimicking experiment. The tools offered can serve as a base for many applications.

## I. INTRODUCTION

Human-robot interaction (HRI) is the study of interactions between humans and robots. HRI is a multidisciplinary field with contributions from human-computer interaction, artificial intelligence, robotics, natural language understanding, design, and social sciences.

A requirement for natural HRI is to endow the robot with the ability to capture, process and accurately and robustly understand human requests. Therefore it is important to analyse the natural ways by which a human can interact and communicate with a robot. A considerable number of robotic systems have been developed in the last decade showing HRI capabilities [5][7].

In recent years, the robotics field has seen the emergence of sophisticated humanoid robots, including Honda Asimo and *NAO*. Due to their human-like morphology, humanoids are well-suited to operate in shared environments with humans. Therefore, they are used by many researchers to investigate fields like navigation in unstructured environments, full body motions and human-robot interaction.

Real-time teleoperation of humanoid robots by detecting and tracking human motion is an active research area. This type of teleoperation can be considered as a particular way of interaction between a person and a robot, allowing an intuitive teleoperational control due to similarities in embodiment between the human master and the robot slave. It is an interesting research topic, the related work is abundant. To mention some, Setapen et al. [15] use motion capture to teleoperate a *NAO* humanoid robot, using inverse kinematic calculations for finding the mapping between motion capture data and robot actuator commands. Matsui et al. [12] use motion capture to measure the motion of both, a humanoid robot and a human, and then adjust the robot's motions to minimise the differences, with the aim of creating more naturalistic movements on the robot. Song et al. [17] use a custom-built wearable motion capture system, consisting of flex sensors and photo detectors. To convert motion capture

data to joint angles, an approximation model is developed by curve fitting of 3rd order polynomials. Koenemann et al. [10] present a system that enables a humanoid robot to imitate complex whole-body motions of humans in real time, ensuring static stability when the motions are executed and capturing the human data with an Xsens MVN motion capture system consisting of inertial sensors attached to the body.

The above mentioned methods are limited in the sense that the human needs to wear different type of sensors in order to interact with the robot. This can be avoided with the Kinect sensor. Moreover, the cost of the equipment is declined. Song et al. [18] propose a new approach for robot control in order to give autonomy to the robot and people's subjective initiative. Suay et al. [20] present a new humanoid robot control and interaction interface that uses depth images and skeletal tracking software to control the navigation, gaze and arm gestures of a humanoid robot. The work described here follows these criteria and develops a Kinect sensor-based gesture teleoperation system.

This paper proposes two new software packages to enrich the teleoperation of *NAO*. The software provides functionality to fully teleoperate the robot in real-time, allowing speech-based guidance, gesture-based teleoperation that includes arm motion control. With the mentioned modules, the robot *NAO* is able to maintain speech-based communication with the user, copy the motion of the arms and walk and rotate in all directions. Experiments in the laboratory as well as public performances have been conducted to evaluate the usefulness of the proposed software modules.

## II. THE ROBOT *NAO*, THE KINECT SENSOR AND SPEECH RECOGNITION WITH ROS

### A. ROS

ROS (Robot Operating System) [13] is a framework for robot software development, and it also provides operating system-like functionality on an heterogeneous computer cluster. The aim of ROS is to be a system that combines some useful elements. These elements are drivers and algorithms (such as navigation algorithms, control algorithms for robotic arms, etc.). The system is based on a modular concept. Modules are named nodes in ROS and nodes communicates via *topics* or *services* following a publisher/subscriber protocol.

### B. ROS for *NAO*

One of the robots for which ROS is available, although in a rather limited way, is *NAO*. *NAO*'s human like shaped body is about 58 cm tall and weights about 4,8 kg. It is built in polycarbonate and abs (a common thermoplastic) materials

\*This work was supported by the Basque Government Research Team Grant (IT313-10), SAIOTEK Project SA- 2013/00334 and the University of the Basque Country UPV/EHU (Grant UFI11/45 (BAILab))

<sup>1</sup>Authors are with Faculty of Informatics, Computer Science and Artificial Intelligence, University Basque Country (UPV/EHU), San Sebastian, igor.rodriguez@ehu.es

that allow better resistance against falls and it has a lithium battery with which it can get an autonomy of 90 minutes approximately. Its heart is composed by a 1.6 GHz Intel Atom processor running Linux. 25 servos enable to control the 25 degrees of freedom of the robot. Regarding to robot motion, *NAO* can move in any direction (omnidirectional walking), it uses a simple dynamic model (linear inverse pendulum) and quadratic programming. It is stabilized using feedback from joint sensors. This makes walking robust and resistant to small disturbances, and torso oscillations in the frontal and lateral planes are absorbed. It can walk on a variety of floor surfaces, such as tiled and wooden floors, and he can transition between these surfaces while walking.

`nao_robot` is the name of the metapackage used to control the robot. It was developed by the University of Freiburg and it is available on the ROS web. ROS can be run on the robot, or remotely, sending appropriate actions and reading information from *NAO* through a wifi. The last option is the alternative used in this work.

### C. Kinect and ROS OpenNI tracker

The Kinect sensor is a horizontal bar connected to a small base with a motorized pivot, and is designed to be positioned lengthwise above or below the video display. The Kinect consists of three different sensors (a RGB camera, a depth sensor and a multiarray microphone) that work together to create the experience of a natural user interface [9].

ROS has several packages that allow to work with the Kinect. The `openni_tracker` package detects when a person gets into the scene captured by the Kinect and tracks the position of his/her head and limbs afterwards. This package makes use of the OpenNI (*Open Natural Interaction* framework).

### D. Speech recognition and ROS *gspeech*

*Google Speech* [19] is an automatic speech recognizer (ASR), freely available, used in several Google applications such as “Google Voice” or “Google Now”. It is a cloud based service in which a user submits audio data using a HTML POST request and receives as reply the ASR output in the form of an n-best list. The user only can customize the number of hypotheses returned by the ASR, specify the language used and enable a filter to remove profanities from the output text. The service returns only the final hypothesis and their corresponding confidence level. Google’s speech engine has the advantages of being speaker independent and that no training is required.

ROS provides a package based on the *Google Speech* tool and dependant on SOX<sup>1</sup>, named `gspeech`. It has a single node that is in charge of:

- 1) Sound capture: when the user starts speaking, the sound captured from the microphone is recorded by SOX and saved into an audio file.

<sup>1</sup>A sound processing program that can convert various formats of audio, apply various effects to those sound files, and also, can play and record audio files on most platforms

- 2) Speech to text conversion: the audio is sent to the Google’s server, which in turn will process it and return the speech content and the confidence value of the recognition.

## III. SPEECH BASED TELEOPERATION IN BASQUE

Verbal communication should be a natural way of human-robot interaction. It is a type of communication that allows the exchange of information with the robot.

To serve a human being, it is necessary to develop an active auditory perception system for the robot that can execute various tasks in everyday environments obeying spoken orders given by a human and answering accordingly. Several systems have been recently developed that permit natural-language human-robot interaction. Foster et al. [6] propose a human-robot dialogue system for the robot JAST, where the user and the robot work together to assemble wooden construction toys on a common workspace, coordinating their actions through speech, gestures, and facial displays. Wang et al. [22] introduce a human-robot speech system for teleoperating a humanoid mobile robot.

A speech based teleoperation interface should provide the user the possibility to teleoperate *NAO* giving predefined orders. The system also should give feedback to the operator when an instruction is not understood and this feedback should also be verbal. Three elements are identified in an architecture for speech-based teleoperation:

- 1) The speech recognition system (SR)
- 2) The text to speech (TTS) system
- 3) The robot control system

Within the available ROS packages for *NAO*, the `nao_speech` package provides the necessary tools for making *NAO* understand and speak in English. But this package is of no use when another language is required, as it is the case. Our *NAO* robot is supposed to interact in **Euskara** (Basque, a minority language spoken in the Basque Country) and thus, a tool adapted to this requirement was needed.

### A. Implementation: The `nao_teleop_speech_eus` package

A new package (`nao_teleop_speech_eus`) has been developed with three different nodes, one for each of the identified elements:

a) `nao_gspeech`: As mentioned before, ROS `gspeech` package gives ASR capabilities to the robot and can be configured for many languages, including Basque. But this package needs some modifications in order to be useful in a real-time teleoperation scenario. These are the introduced changes:

- 1) When the native `gspeech` runs the Google’s speech service, it is executed only once, i.e., when the user starts speaking, the audio is captured and sent to Google. There it is analysed and the text “corresponding” to the received audio is returned with a confidence level; then, the program ends. It could be tedious for the user to run the speech recognition node each time she/he wants to order something to the robot, or each

time she/he receives an error message. Hence, the new node now runs iteratively avoiding the problem of having to launch the node each time the user wants to talk.

- 2) When the Google's speech service does not recognize the spoken words, it returns an error message and then the node is forced to quit. Now, error messages received from Google's speech service are specially treated. If an error message is received, `nao_gspeech` publishes a `Repeat` message in the `google_speech` topic to advertise the user that his spoken words are not being recognized.
- 3) The original `gspeech` node only prints the response received, it does not publish any messages or services, so it can not communicate with other nodes. After the modifications, the confidence level of the hypothesis received from *Google Speech* is processed and, if it is lower than a predefined threshold (0.15 for the performed experiments), the response is declined and treated as an error message.

b) `nao_tts_eus`: This is the node in charge of converting the text into speech using the *AhoTTS* tool [11]. That system, developed by the Aholab group in the University of the Basque Country, is a modular text to speech synthesis system with multithread and multilingual architecture. It has been developed for both, Euskara and Spanish languages. The TTS is structured into two main blocks: the linguistic processing module and the synthesis engine. The first one generates a list of sounds, according to the Basque SAMPA code [14], which consists of the phonetic transcription of the expanded text, together with prosodic information for each sound. The synthesis engine gets this information to produce the appropriate sounds, by selecting units and then concatenating them and post-processing the result to reduce the distortion that appears due to the concatenation process. This tool is required for communicating in Basque Language, but it would not be required for English interlocution.

`nao_tts_eus` has a subscriber that receives messages from `nao_teleop_speech` (see below) in the `text_speech` topic. When a text message is received, this node converts it into speech (an audio file) and plays the audio over *NAO's* speakers.

c) `nao_teleop_speech`: This node allows the user to control *NAO's* movements using several voice commands. The operator, situated in the teleoperation cab (the place where the remote PC is located), gives orders to the robot using a microphone. The robot is able to perform these movements: *Stand up, Sit down, Move forward, Move backward, Move left, Move right, Turn left, Turn right* and *Stop*

As we previously said, commands are given in Euskara. With the intention to communicate with the robot in a more natural way, the user has more than one choice for each possible command. That is, if the operator wants the robot to stand up, he can say: "Altxatu", "Tente jarri", "Zutik jarri", etc. Therefore a dictionary of some predefined words has been created, including several synonymous for each command. When the user gives a voice command (it can be a

long sentence), the voice is converted to text, and processed afterwards. The system tries to find matches between the dictionary and the text received from Google's speech service. If a match is found, the robot performs the movement corresponding to the received command, otherwise the robot says that it could not understand the order and asks the user to repeat it.

`nao_teleop_speech` is the node in charge of receiving messages from `nao_gspeech`, finding any matches in the predefined commands dictionary and deciding what is the action that *NAO* must perform. It has a subscriber to receive messages that `nao_gspeech` publishes on the `google_speech` topic, and two publishers; one to set *NAO's* walking velocity according to the given command, and another one to publish the text messages that *NAO* has to say.

In order to test the speech capabilities in a HRI context, we integrated the ASR and TTS modules in our Bertsobot project [1]. The Bertsobot project was showed to the general public in a live performance entitled "Minstrel robot: science or fiction"<sup>2</sup>, in which *NAO* robot showed his verse-impromvisation and speech-based communication capabilities.

It must be said that the system, and as a consequence, the speech based teleoperation works when the google's voice server gives the correct answer. The drawback is that Google can disable the service without notice.

#### IV. GESTURE BASED TELEOPERATION AND ARM MOVEMENT

We want to control the robot's walking motion and also manipulate his arms based on the movements the operator makes. Hence, the gesture teleoperation proposed in this work uses all body gestures. Arm movements are used to control the robot arms, i.e. *NAO* replicates the operator's arm movements. Besides, the operator's position (of her/his body) defines the action she/he wants to perform, like move forward or turn left.

##### A. Gesture-based teleoperation

The *NAO* teleoperation process has two stages; the operator's calibration stage and the robot control stage. First of all, the operator must perform the calibration pose in front of the Kinect view, this is a necessary step because it verifies that the Kinect is seeing a person. The calibration process is done by `openni_tracker`. When the calibration process is successful, the system is ready to receive gesture commands.

The commands that the robot understands are: *Stand up, Crouch, Stop, Move forward/backward, Move left/right* and *Turn left/right*. Figure 1 shows the gestures required to the user in order to command the robot.

Those gestures must be performed from a particular position, since the system is configured for specific user positions and distances (see the black marks on the floor in figure 1). The teleoperation process starts with the operator in crouch

<sup>2</sup>[http://www.alhondigabilbao.com/web/guest/detalle-evento/-/journal\\_content/56\\_INSTANCE\\_aJV5/10140/4092781?last-page=/programacion/badu-bada](http://www.alhondigabilbao.com/web/guest/detalle-evento/-/journal_content/56_INSTANCE_aJV5/10140/4092781?last-page=/programacion/badu-bada)



Fig. 1. Gesture teleoperation commands

position, because this is the position *NAO* rests when it is not being operated. When the operator stands up, the robot stands up too and it is ready to receive moving orders.

### B. Arm control

The goal of this process is to give the operator the option to move the robot arms remotely. This can be very useful if the robot is involved in a task where it must carry objects.

Human arms have 7 degrees of freedom (DOF): three at the shoulder and wrist, and one at the elbow (yaw). On the contrary, *NAO*'s arms have five DOFs, two at the shoulder (pitch and yaw) and elbow (yaw and roll), and one at the wrist (yaw) (figure 2). Thereby, the movement configurations of human and robot arms differ. For sake of simplicity, we will limit the robot arm movement to raise and extend the arms, imitating the operator's arm motions.

Imitation starts with the perception of the demonstrator; the operator must perform the calibration pose and when it succeeds, the system is ready to perform the imitation process.

For the purpose of imitation, first human movement data have to be acquired, then data are suitably transformed to *NAO*'s coordinate space, and finally, the gesture is executed by *NAO*. We will name this problem as the correspondence problem. This is the main problem of this task.

In order to solve the correspondence problem, the arm is modelled as a stick. The data are acquired by using *openni\_tracker* and provide Cartesian coordinates of twenty joints on the human body. The joints tracked by *openni\_tracker* are: *Head*, *Neck*, *Torso*, *LeftShoulder*, *LeftElbow*, *LeftHand*, *RightShoulder*, *RightElbow*, *RightHand*, *LeftHip*, *LeftKnee*, *LeftFoot*, *RightHip*, *RightKnee* and *RightFoot*. The Kinect treats itself as the origin when providing the Cartesian coordinates, but *NAO* has its own reference system with a different origin. For this reason, Kinect's joint information must be translated.

To transform the Cartesian coordinates obtained from the Kinect into *NAO*'s coordinate space a joint control approach was employed. In this approach, we calculate the human joint angles at shoulders and elbows from Cartesian coordinates using inverse kinematics (as described below), and command the robot to set the arms' position at the calculated angles.

**Transformation from Kinect's coordinate space to *NAO*'s space:** We will focus the explanation on the left arm. The analysis of the right arm is similar and it will be omitted here.

*NAO*'s left arm has four joints (see figure 2): *LShoulderRoll*, *LShoulderPitch*, *LElbowRoll*, *LElbowYaw* and *LWristYaw*. *LElbowYaw* and *LWristYaw* joints have not been taken into consideration for the joint control proposed in this work, because the *openni\_tracker* package can not detect the operator's arms yaw motion.

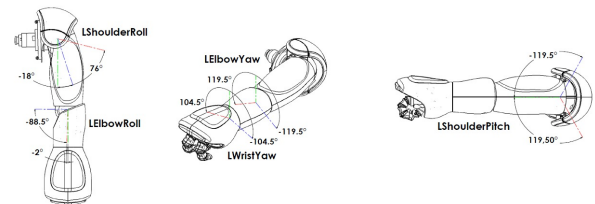


Fig. 2. Left arm motion

In order to calculate the  $LShoulderRoll$  motion angle ( $\alpha_{ShoulderRoll}$ ) we use the dot product between  $LRSoulder$  and  $LShoulderElbow$  vectors, where  $LRSoulder$  is the vector between the right and left shoulder joints, and  $LShoulderElbow$  is the vector between the left elbow and left shoulder joints.

Since  $LRSoulder$  and  $LShoulderElbow$  vectors are known, we can obtain the angle between these two vectors using the dot product. Note that before applying the equation,  $LRSoulder$  and  $LShoulderElbow$  vectors must be normalized. The  $\alpha_{ShoulderRoll}$  angle can be calculated as follows:

$$\alpha_{ShoulderRoll} = \arccos(LRSoulder \cdot LElbowShoulder) \quad (1)$$

The  $\alpha_{ShoulderRoll}$  angle is calculated in the Kinect's coordinate space, therefore, it must be transformed into NAO's coordinate space by rotating it  $\frac{-\pi}{2}$  radians. Figure 3 shows both coordinate frames and the transformation needed.

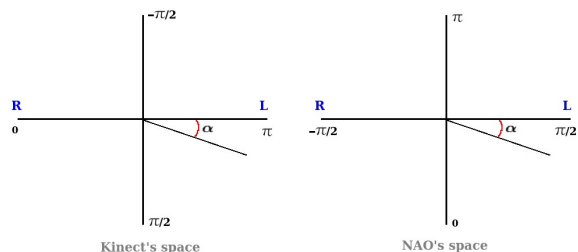


Fig. 3. Left shoulder roll motion in Kinect and NAO spaces

$LElbowRoll$  motion angle is calculated in the same way as  $LShoulderRoll$  motion angle. But now, to calculate the  $\alpha_{ElbowRoll}$  angle the vectors we need to consider are  $LShoulderElbow$  and  $LHandElbow$ , where  $LShoulderElbow$  is the vector between the left shoulder joint and the elbow and  $LHandElbow$  is the vector between the left hand joint and the elbow.

Again,  $\alpha_{ElbowRoll}$  angle must be transformed to NAO's space, in this case by rotating it  $-\pi$  radians.

And finally, the  $LShoulderPitch$  motion angle, which is calculated taking into account only the height the  $LShoulderElbow$  vector takes with respect to  $z$  axis. When the arm is extended making an angle of 90 degrees with the torso, it is considered on the  $z=0$  axis. Therefore, if you raise the arm from this position the angle will be positive, and if you lower it, negative (see figure 4).

After normalizing the  $LShoulderElbow$  vector, the  $\alpha_{ShoulderPitch}$  angle for  $LShoulderPitch$  motion can be defined as:

$$\sin(\alpha_{ShoulderPitch}) = \frac{\|A\|}{\|LShoulderElbow\|} = \frac{\|A\|}{1} \quad (2)$$

$$\|A\| = z_{LShoulderElbow} \quad (\text{by definition}) \quad (3)$$

$$\alpha_{ShoulderPitch} = \arcsin(z_{LShoulderElbow}) \quad (4)$$

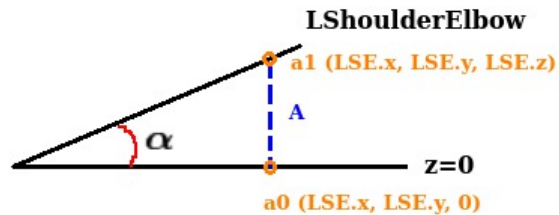


Fig. 4. Left shoulder pitch motion in the Kinect space

where  $z_{LShoulderElbow}$  is the  $Z$  coordinate of  $LShoulderElbow$ , represented as  $LSE.z$  in figure 4. To transform the  $\alpha_{ShoulderPitch}$  angle to NAO's space it must be rotated  $2\pi - \alpha_{ShoulderPitch}$  radians.

### C. Implementation: The nao\_teleop\_gesture package

A new package named `nao_teleop_gesture` has been developed in order to achieve the gesture-based teleoperation system. This new package contains two nodes:

- `nao_motion_control`: it has to perform the following two main tasks:
  - 1) Receive messages published by the `openni_tracker` package.
  - 2) Publish NAO's walking velocities.

So for, the `nao_motion_control` node contains a subscriber that receives messages published by `openni_tracker` in the `skeleton` topic. When a message is received, the operator's position is analyzed. And according to that position the robot performs the corresponding walking motion. For example, if the message received indicates that the user took a step forward, the robot starts walking forward. Although each gesture indicates a movement order, different gestures can be combined. i.e. combining move forward and turn left gestures the robot can move forward rotating to the left at the same time. Only walking action movements (forward, backward, left, right) with rotational motions can be combined.

On the other hand, the `nao_motion_control` node has a publisher that publishes the omnidirectional velocity ( $x$ ,  $y$ , and  $\theta$ ) for the walking engine. The velocity with which the robot moves has been set to a constant value. If the walking velocity is too high the robot starts to swing. Thus, the linear velocity and the angular velocities have been assigned low values. The walking velocity of the robot is published in the `cmd_vel` topic.

- `nao_arm_control`: This node is in charge of sending to the robot the necessary motion commands to replicate the operator's arms motion. The node performs tasks by:
  - 1) Receiving the messages published by the `openni_tracker` package.



2) Publishing *NAO*'s joint angles with speed.

Therefore, `nao_arm_control` is subscribed to the `skeleton` topic in order to receive the operator's skeleton messages published by `openni_tracker`. On the other hand, the `nao_arm_control` node has a publisher that publishes the joint angles with speed in the `joint_angles` topic, which allows the communication with the `nao_controller` node. The *NAO*'s joints motion speed is set to a constant value appropriate for the robot to mimic the operator arms motion in "real" time.

The robot imitates human actions in real-time with a slight delay of less than 30 milliseconds. This delay is approximately the time the system needs to capture the operator's arms motion, calculate the angles that make up the operator's arm joints, and send motion commands to the robot via Wi-Fi.

It must be mentioned that the native `nao_robot` package has been enriched to provide ultrasound information from *NAO*'s chest and `nao_teleop_gesture` now can use that information, to alert the user when *NAO* gets too close to obstacles in front.

### V. SYSTEM EVALUATION: MIMICKING BEHAVIOR

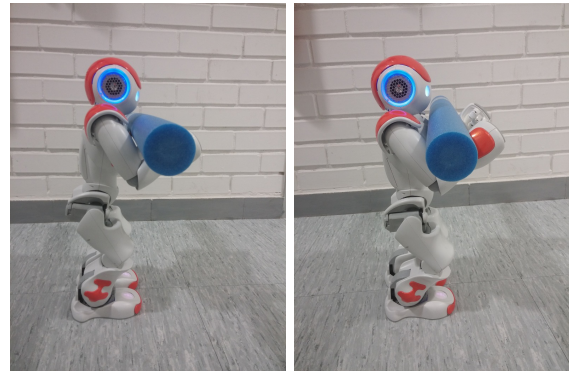
Imitation is an important way of skill transfer in biological agents. Many animals imitate their parents in order to learn how to survive. It is also a way of social interaction. A sociable robot must have the capability to imitate the agents around it. In a human society, people generally teach new skills to other people by demonstration. We do not learn to dance by programming, instead we see other dancers and try to imitate them. Hence, our artificial partners should be able to learn from us by watching what we do.

In order to evaluate the developed system we propose a set of experiments based on imitation. The speech-based teleoperation was not included in the experiments due to some problems during the transition to version 2 of the *Google Speech Server*.

The mimicking behavior consists of the arm control behavior together with the teleoperation behavior. The walking behavior of the robot has been modified for those cases when the robot has something on its arms. The center of gravity (COG) of the walking robot has been lowered and backwarded so that the COG is maintained within the support polygon (see figure5).

The process runs as follows. *NAO* starts in crouching position and when the operator enters the Kinect's view, the calibration process starts. *NAO* tells the operator that the calibration ended successfully saying "Kinect control enabled" in Basque and then, the operator can control the robot with his/her body.

A GUI has been created (with existing *rqt* plugins) in order to help the operator to know the system state. It can be divided into two main parts (figure 6). The top side is composed by the *Topic Monitor* and the *Rviz* interface. The *Topic Monitor* shows all the topics and messages sent by the nodes that are in execution. *Rviz* shows the *NAO* 3D



(a) Original (b) Modified

Fig. 5. Modified walking position

model moving in real-time. The bottom side shows visual information from the cameras; *Image View* shows the image received from *NAO*'s top camera and the right window shows the image captured by the Kinect together with the skeleton of the tracked body.

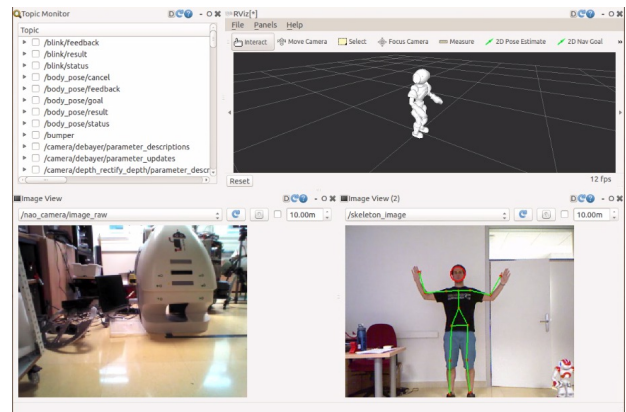


Fig. 6. Teleoperation display

Two experiments were defined. Those experiments involved several people that should give qualitative measures of the system performance by means of a questionnaire that participants completed after carrying out each experiment. Experiments were performed until each participant achieved the aim of the experiment at least once. To run the experiments a package named `nao_imitation` has been created with just a launch file that run both nodes in the `nao_teleop_gesture` package and the *rqt* GUI interface. All the code is available at [github.com/rsait/NAO-ROS](https://github.com/rsait/NAO-ROS).

**Experiment #1:** In this first experiment, the operator has to drive the robot in the circuit formed by a series of obstacles using a predefined sequence of movements that included all the movements available in the system: walking, lateral displacement and rotations. Figure 7 shows the circuit proposed, together with the movements that must be performed.

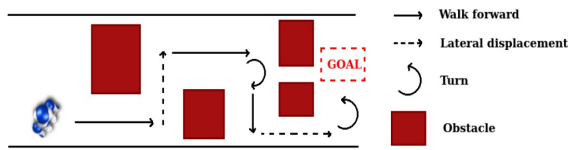


Fig. 7. Test environment for experiment #1

The aim of this experiment was to test if the control of the robot could be carried out with accuracy in reduced spaces, and the adequateness of the robot's view about the environment for the operator to properly teleoperate the robot. It must be emphasized that during the experiments users were not in the same environment as the robot, but in a different lab with a monitor where the GUI displayed the scene view of the robot. Moreover, the only knowledge about the scenario given to the user was the schema shown in figure 7.

Table I shows the results of questionnaires completed by the participants.

User	Goal reached	Attempts	Precision of movements	Difficulty
1	Yes	2	3	4
2	Yes	2	3	3
3	Yes	2	4	4
4	Yes	3	5	3
5	Yes	1	5	2
Total	5/5	2	4/5	3.2/5

TABLE I  
RESULTS OF THE FIRST EXPERIMENT

As shown in table I, all participants successfully reached the goal. Most participants needed more than one attempt to complete the experiment. Some of them failed in the first attempt because the perception of the environment was not good and they were not able to avoid all the obstacles. More specifically, the lack of side view difficults the guidance of the robot.

On the other hand, despite the difficulty of the experiment, most participants believed that the selection of gestures is correct (natural) and the movements are quite precise. We can conclude that the system requires a short period of training for the operator to get used to the distances.

**Experiment #2:** In the second experiment, the operator must operate the robot to transport an object from one place to another. To do it, the operator must drive the robot towards the person who will give it the object that has to take in its arms. Once the robot has taken the object, the operator must send the robot towards the other person who will collect the object. The aim of the experiment was to test the movements of the robot arms while performing a simple task in cooperation with humans. Figure 8 shows the test

environment of this experiment.

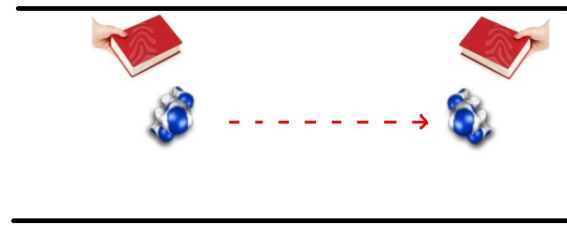


Fig. 8. Test environment for experiment #2

Table II shows the results of the second experiment obtained in the questionnaire completed by the participants.

User	Goal reached	Attempts	Precision of movements	Precision of arms	Difficulty
1	Yes	2	5	5	2
2	Yes	4	5	2	3
3	Yes	2	4	5	2
4	Yes	1	5	5	1
5	Yes	1	4	5	1
Total	5/5	2	4.6/5	4.4/5	1.8/5

TABLE II  
RESULTS OF THE SECOND EXPERIMENT

Even though according to table II all participants successfully reached the goal, some needed more than one attempt to complete the aim of the experiment; two of them failed at least once because the robot lost balance when walking with the arms raised, and the other one failed because the Kinect did not detect well (for unknown reasons) the participant's left arm movements.

## VI. CONCLUSIONS AND FURTHER WORK

The main contribution of this paper is a set of freely available ROS packages that serve for teleoperating a humanoid robot. A speech based teleoperation package has been developed for Basque, but it can be easily available to other languages as long as *Google Speech Service* provides speech recognition for it. Right now it supports about 35 languages<sup>3</sup>. A new node has been created taking as reference the native *gspeech* ROS package that allows to command the robot verbally repeatedly in spite of network errors or misunderstood commands. Even though the used TTS is not as general as the ASR system, the node can be adapted by changing the code to use a different TTS system's API.

On the other hand, the gesture teleoperation system together with the user interface allows the robot to imitate arm movements and act accordingly to predefined body movements made by the operator at the same time. We are now working on some improvements such as letting the operator to dynamically adjust the velocity of the walking commands (till now, the velocities were constant and fixed

<sup>3</sup>[http://en.wikipedia.org/wiki/Google\\_Voice\\_Search](http://en.wikipedia.org/wiki/Google_Voice_Search)



experimentally). New movements are also being studied such as that of the head. Allowing the user to control the head position would give the user the chance to take lateral views without moving the robot, relieving the effect of the narrow view of the camera (video of the ongoing progress at <http://www.sc.ehu.es/ccwrobot>).

Something that remains to be done in the arm control is the yaw movement of *NAO*'s elbows and hands. *NAO* can rotate its arms from the elbows, also its hands from the wrists. A future solution to this problem could be based on the work done by Cole, Grimmes and Rao [3]. They propose a system able to learn full-body motions from monocular vision. In order to detect body motions they use different colors to identify different body parts. Considering the idea of identifying body parts with different colors, the top and bottom sides of the arms and hands could be marked with different colors, enabling to distinguish the palm and the back of the hands.

With respect to the usefulness of the system, we are offering a tool that can have many applications beyond pure teleoperation. Within the area of socially assistive robotics, i.e. robots that assist through social interaction [4], many authors agree that humanoid robots can help assisting patients with disabilities[16], or in pediatric therapy [8][2] because children tend to accept robots in a more natural way than adults. In [21] authors present an experience using the *Nao* humanoid robot in a role of a physiotherapist for rehabilitation and prevention of scoliosis in children. Children are motivated to replicate robots' movements as a therapy. The developed software could help the therapists without programming skills to record the exercises on the robot. Our next step will be to provide a tool for recording or learning sequences of poses from the mimicked movements.

## REFERENCES

- [1] A. Astigarraga, M. Agirrezabal, E. Lazkano, E. Jauregi, and B. Sierra. Bertsobot: the first minstrel robot. In *Human System Interaction*, pages 129–136, 2013.
- [2] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayhuatl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayov, G. Athanasopoulos, V. Enescu, R. Looije, M. Neerinx, Y. Demiris, R. Ros-Espinoza, A. Beck, L. Caamero, A. Hille, M. Lewis, I. Baroni, M. Nalin, P. Cosi, G. Paci, F. Tesser, G. Sommavilla, and R. Humbert. Multimodal child-robot interaction: Building social bonds. *Human-Robot Interaction*, 1(2):33–53, 2012.
- [3] J. C. Cole, D. B. Grimes, and R. P.N. Rao. Learning full-body motions from monocular vision: Dynamic imitation in a humanoid robot. In *Intelligent Robots and Systems (IROS)*, pages 240–246, 2007.
- [4] D. Feil-Seifer and M-J. Matarc. Defining socially assistive robotics. In *9th International Conference on Rehabilitation Robotics*, pages 465–468, 2005.
- [5] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.
- [6] M. E. Foster, M. Giuliani, A. Isard, C. Matheson, J. Oberlander, and A. Knoll. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *IJCAI*, pages 1818–1823, 2009.
- [7] M. A. Goodrich and A. C. Schultz. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 1(3):203–275, 2007.
- [8] Ayanna M. Howard. Robots learn to play: Robots emerging role in pediatric therapy. In *26th International Florida Artificial Intelligence Research Society Conference*, 2013.
- [9] S. Kean, J. Hall, and P. Perry. *Meet the Kinect: An introduction to programming natural user interfaces*. Technology in action, 2011.
- [10] J. Koenemann and M. Bennewitz. Whole-body imitation of human motions with a nao humanoid. In *HRI*, pages 425–426, 2012.
- [11] I. Leturia, A. Del Pozo, K. Arrieta, U. Iturraspe, K. Sarasola, A. Ilaraza, E. Navas, and I. Odriozola. Development and evaluation of anhitz, a prototype of a basque-speaking virtual 3d expert on science and technology. In *Computer Science and Information Technology, 2009. IMCSIT'09*, pages 235–242, 2009.
- [12] D. Matsui, T. Minato, K. F. MacDorman, and H. Ishiguro. Generating natural motion in an android by mapping human motion. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3301–3308. IEEE, 2005.
- [13] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Y. Ng. Ros: an open-source robot operating system. In *Open-Source Software workshop of the International Conference on Robotics and Automation (ICRA)*, 2009.
- [14] SAMPA. <http://bips.bi.ehu.es/sampa.html>.
- [15] A. Setapen, M. Quinlan, and P. Stone. Beyond teleoperation: Exploiting human motor skills with marionet. In *AAMAS 2010 Workshop on Agents Learning Interactively from Human Teachers (ALIHT)*, 2010.
- [16] M. Simonov and G. Delconte. Assessment of rehabilitative exercises by humanoid robot. In *7th Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 331–334, 2013.
- [17] H. Song, D. Kim, M. Park, and J. Park. Tele-operation between human and robot arm using wearable electronic device. In *17th IFAC World Congress*, pages 2430–2435. IFAC, 2008.
- [18] W. Song, X. Guo, F. Jiang, S. Yang, G. Jiang, and Y. Shi. Teleoperation humanoid robot control system based on kinect sensor. In *4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 2, pages 264–267. IEEE, 2012.
- [19] Google Speech. <https://www.google.com/intl/es/chrome/demos/speech.html>.
- [20] H. B. Suay and S. Chernova. Humanoid robot control using depth camera. In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 401–401. IEEE, 2011.
- [21] M. Vircikova and P. Sincak. Experience with the children-humanoid interaction in rehabilitation therapy for spinal disorders. In J. H. Kim, E. T. Matson, and H. Myung an P. Xu, editors, *Robot Intelligence Technology and Applications*. Springer, 2013.
- [22] B. Wang, Z. Li, and N. Ding. Speech control of a teleoperated mobile humanoid robot. In *IEEE international conference on automation and logistics*, pages 339–344. IEEE, 2011.

## 11.4 NAO Robot as Rehabilitation Assistant in a Kinect Controlled System

**Title:** NAO Robot as Rehabilitation Assistant in a Kinect Controlled System

**Authors:** I. Rodriguez, A. Aguado, O. Parra, E. Lazkano and B. Sierra

**Conference:** International Conference on Neurorehabilitation (ICNR)

**Publisher:** Springer

**DOI:** DOI 10.1007/978-3-319-46669-9\_70

**Year:** 2017

# NAO Robot as Rehabilitation Assistant in a Kinect Controlled System

I. Rodríguez, A. Aguado, O. Parra, E. Lazkano and B. Sierra

**Abstract** In this paper NAO robot is presented as a Home Rehabilitation assistant; Machine Learning is used to classify the data provided by a Kinect RGB-D sensor in order to obtain a Home Exercise Monitoring System which aims at helping physicians controlling patient at home rehabilitation.

## 1 Introduction

Rehabilitation robotics is a field of research dedicated to understanding and augmenting rehabilitation through the application of robotic devices, and includes development of robotic devices tailored for assisting different sensorimotor functions (e.g. arm, hand, leg, ankle); here, robots are used mainly as therapy aids instead of assistive devices.

Human position estimation is an important fact which needs to be considered when rehabilitation applications are developed. In this paper a new approach is presented, which combines, on the one hand, the Kinect sensor provided data, and on the other, different paradigms of the Machine Learning area, both supervised and unsupervised, to achieve a good classification of the perceived human position. Once the positions are defined, a rehabilitation system which aims at achieving good starting and ending positions to several movements is developed.

## 2 Materials and Methods

The rehabilitation system proposed in this paper, which aims at maintaining the people adherence when rehabilitation exercises are to be done at home; to do that, on

---

I. Rodríguez (✉) · A. Aguado · O. Parra · E. Lazkano · B. Sierra  
Robotics and Autonomous Systems Research Group, Computer Science  
and Artificial Intelligence Department, University of the Basque Country, Leioa, Spain  
e-mail: igor.rodriguez@ehu.eus

© Springer International Publishing AG 2017  
J. Ibáñez et al. (eds.), *Converging Clinical and Engineering Research on Neurorehabilitation II*, Biosystems & Biorobotics 15,  
DOI 10.1007/978-3-319-46669-9\_70

419

the one side, we use a NAO robot to teach the patient the rehabilitation activities to be performed, and on the other side, we use an RGB-D sensor to perceive the human body positions and classify them using Machine Learning classifiers. Movements are monitored automatically, and the system evaluates the movement quality. Improvements from one day to other can also be automatically detected. The rehabilitation system has been developed using ROS (Robot Operating System) framework.

## 2.1 NAO

NAO is an autonomous programmable humanoid robot developed by Aldebaran Robotics [3]. The main goal of this robot was to provide a hardware and software platform that will allow progress in that research area at a reasonable cost.

NAO's human-like shaped body is about 58 cm tall, weights about 4.8 kg and it can get an autonomy of 90 min approximately (see Fig. 1). 25 servos enable to control the 25 degrees of freedom of the robot and it can move in any direction (omnidirectional walking).

Many different applications have been developed for NAO since its birth: it can play football, it is able to recognize objects, faces or voices, collaborate with other NAO robots to load objects, obey orders, write, perform group choreography, play a xylophone, help in the kitchen, and many other things.

## 2.2 Kinect Sensor

The Kinect sensor [2], created by Microsoft to control and interact with their console/computer using gestures and spoken commands, consists of an RGB camera, a depth sensor (IR laser projector combined with a monochrome CMOS sensor) and a multiarray microphone running proprietary software which provide full-body 3D motion capture (as shown in Fig. 2), facial recognition and voice recognition capabilities.

**Fig. 1** NAO doing exercise



**Fig. 2** Depth image provided by Kinect



### 2.3 ROS and OpenNI

ROS (Robot Operating System) [4] is a framework for robot software development; the aim of ROS is to be a system that combines some useful drivers and algorithms (such as navigation algorithms, control algorithms for robotic arms, etc.). The system is based on a modular concept, which means that we can have different modules performing different tasks, and also can communicate each other.

ROS has several packages that allow to work with the Kinect. The *openni\_tracker* package detects when a person gets into the scene captured by the Kinect and tracks the position of his/her head and limbs afterwards. Figure 3 shows the users' head and limbs detection

**Fig. 3** Users's body tracking using OpenNI framework



To perform the pose classification several Machine Learning Standard classifiers have been used; the best result has been obtained by Random Forest paradigm [1].

## 2.4 Used Data

Different human poses were defined for classification tasks. The total number of defined poses is 16: these are divided into three different groups. The first group (poses 1–11) are arm related poses. The second group (12–15) are global poses, related to the whole skeleton. The last pose (16) is any other pose which will not be recognized. In our Machine Learning approach, we will train a classification model for each group of poses.

(1) psi	(7) left-arm-up	(13) sit
(2) arms-wide	(8) right-arm-right	(14) bend
(3) arms-front	(9) right-arm-front	(15) lie
(4) arms-up	(10) right-arm-up	(16) none
(5) left-arm-left	(11) hello	
(6) left-arm-front	(12) stand	

The features for the classification models are obtained from the skeleton information provided by the OpenNI library, while OpenNI processes the Kinect sensor data. OpenNI tracks and provides 15 skeleton points.

## 3 Results

Experiments have been made with a different data set for each group of poses. The raw dataset has at least 200 known instances for each pose. The first final dataset (arms) has 2666 instances and 18 features, and the second dataset (global poses) has 2551 instances and 42 features. Tenfold cross-validation has been used for all the test results. Table 1 shows the classification results of arms and body poses performed in the experiments.

**Table 1** Arm and body pose classification results

Classifier	Arm pose	Global pose
	% Correctly classified	% Correctly classified
IB1	99.66	97.96
Naive Bayes	99.62	82.24
C4.5	99.62	97.41
Random Forest	99.85	99.41

## 4 Discussion

The presented system uses Machine Learning paradigms to classify human poses; NAO robot is used to motivate the users to perform the rehabilitation exercises at home, and a Kinect sensor to measure the quality of the movements. A report of the daily and weekly exercises performed by the patients can be automatically obtained and sent to the physicians. In this way, the patients adherence—medical term to indicate patients perform properly the given homework—can be monitored and therefore, its usefulness increases.

## 5 Conclusions

The presented system aims to monitor home rehabilitation exercises. NAO is used as exercise guide, and the Kinect is used to track the movement and to calibrate the quality of the positions.

Obtained results are very promising; nevertheless, as future work, the system needs to be extended in order to be able to perform a better calibration process in an automatic way, to adapt the classification to the patient's characteristics. New classifiers are to be tested to deal with the data provided by Kinect, aiming at obtaining a better accuracy in a more exhaustive scenario where more positions are considered.

Automatic reports of the daily and weekly exercise performed at home are very interesting to the doctors following the rehabilitation, and this is the next step to be done in the system as further work.

**Acknowledgments** This work has been partially supported by the Basque Government (IT900-16) and the Spanish Ministry of Economy and Competitiveness MINECO (TIN2015-64395-R).

## References

1. L. Breiman, Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
2. Kinect, Microsoft Kinect XBOX. <https://support.xbox.com/en-US/xbox-360/accessories/kinect-sensor-components>
3. NAO, Aldebaran Robotics. <https://www.aldebaran.com/en/humanoidrobot/nao-robot>
4. M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A.Y. Ng, ROS: an open-source robot operating system, in *ICRA Workshop on Open Source Software*, vol. 3 (2009), p. 5

# Chapter 12

## Publications related to Part III

### 12.1 Singing minstrel robots, a means for improving social behaviors

**Title:** Singing minstrel robots, a means for improving social behaviors

**Authors:** I. Rodriguez, A. Astigarraga, T. Ruiz, E. Lazkano

**Conference:** International Conference on Robotics and Automation (ICRA)

**Publisher:** IEEE

**DOI:** 10.1109/ICRA.2016.7487454

**Year:** 2016



# Singing minstrel robots, a means for improving social behaviors

Igor Rodriguez<sup>1</sup>, Aitzol Astigarraga<sup>1</sup>, Txelo Ruiz<sup>2</sup> and Elena Lazkano<sup>1</sup>

**Abstract**—*Bertsolaritza*, Basque improvised contest poetry, offers another sphere to develop robot body language and robot communication capabilities, that shares some similarities with theatrical performances. It is also a new area to work on social robotics. The work presented in this paper makes some steps forward in designing and implementing the set of behaviors the robots need to show in the stage to increase, on the one hand robot autonomy and on the other hand, credibility and sociability.

## I. INTRODUCTION

Basque, euskara, is the language of the inhabitants of the Basque Country. And *bertsolaritza*, Basque improvised contest poetry, is one of the manifestations of traditional Basque culture that is still very much alive (see Fig. 1). Events and competitions are very common in which improvised verses, *bertso*-s, are composed. In such performances, one or more verse-makers, named *bertsolari*-s, produce impromptu compositions about topics or prompts which are given to them by an MC (theme-prompter). Then, the verse-maker takes a few seconds, usually less than one minute, to compose and sing a poem along the pattern of a prescribed verse-form that also involves a rhyme scheme. Melodies are chosen from among hundreds of tunes. Xabier Amuriza, a famous verse-maker that modernized and contributed to the spread out of the *bertsolaritza* culture, defined *bertsolaritza* in a verse as:

<i>Neurriz eta errimaz kantatzea hitza horra hor zer kirol mota den bertsolaritza.</i>	<b>Through meter and rhyme to sing the word that is what kind of sport <i>bertsolaritza</i> is.</b>
--	---



Fig. 1. 2009 national championship

Computer-based poetry has been paid attention to in the research community for the last years (see [8] and [21])

<sup>1</sup>Department of Computer Science and Artificial Intelligence, Faculty of Informatics, University of Basque Country (UPV/EHU), 20018 Donostia igor.rodriguez@ehu.eus

<sup>2</sup>Department of Computer Architecture and Technology, Faculty of Informatics, University of Basque Country (UPV/EHU), 20018 Donostia txelo.ruiz@ehu.eus

for a review), but among the several differences that exist between poetry and *bertsolaritza*, mainly the later belongs to the oral genre, and the public performance is extremely important. Therefore, it is not enough the development of an automatic verse generation system, the created poem has to be part of a performance. Thus, a real body that interacts with the public and sings the improvised verse with a proper melody is needed. The interaction with the robot should be speech-based; thus, on the one hand the system should be able to receive the verse requirements to generate the most appropriate verse according to the given instructions and to sing it with the proper melody. On the other hand, the robot must show the same degree of expressiveness Basque troubadours, *bertsolari*-s, do. And all those tasks must be accomplished concurrently in an extemporaneous performance.

We believe that the *BertsoBot* project provides a huge opportunity to join together the capabilities of autonomous robots to sense their environment and interact with it, and the natural language processing tools devoted to automatic verse generation.

## II. RELATED WORK

Human-robot interaction (HRI) is the study of interactions between humans and robots. HRI is a multidisciplinary field with contributions from human-computer interaction, artificial intelligence, robotics, natural language understanding, design, and social sciences. A considerable number of robotic systems has been developed in the last decade showing HRI capabilities [6][9].

But social robots are beyond HRI. According to Breazeal [3], sociable robots are socially intelligent robots in a human like way, and interaction with them is like interacting with persons.

Verbal communication is a natural way of interaction among humans. However, non-verbal expression is key to understand sociability [14]. A bunch of work focuses on facial expressiveness [10][13]. Breazeal's Kismet robotic head represents itself a milestone as how the human voice affects expressiveness. Besides, the advent of humanoid robots has launched researchers to investigate and develop body language expression in robots. Aldebaran's *Pepper* [22] is surely the commercial robot with the highest bodily expression capabilities right now. It has no legs, but it uses its waist and arms to show human like expression while talking.

Robot performances have shown to be a window display for disclosing the state of the art of social robots to the general public, and as such, to measure social acceptance of robots. Although everything is rehearsed beforehand, theater

offers an invaluable sphere to research and develop social behaviors in robots, to work and extend the expression of emotions and the natural communication among humans and robots [18][5]. No need to mention that the term Robot was first used in a play entitled RUR (*Rossum's Universal Robots*) [4]. A review of robot performances can be found in [20]. Little by little robots are bursting into theaters motivated by researchers as a means, but also by artists [17].

However, social robots require to be autonomous. Synthetic replicates are mostly teleoperated or preprogrammed robots; the degree of autonomy shown by performer robots is still far from showing human like behavior (see [19] for a categorization and classification of robots acting in theaters). In our opinion, *Bertsolaritza* offers another sphere to develop robot body language and robot communication capabilities, that shares some similarities with theatrical performances, but also a new area to work on social robotics.

Joxerra Gartzia [7] enumerates the communication act in 5 steps: “*Inventio* (create the message), *dispositio* (give the message the correct form, think how to transmit that message), *elocutio* (how to say the previously prepared message, manage space and time), *memoria* (keep in mind previous work) and *actio* (the action itself)”. Acting needs *elocutio*, *memoria* and *actio*. However, *bertsolaritza* needs to go through the five steps, *inventio* and *dispositio* are mandatory. We'll try to enumerate the main differences between theatrical performances and *Bertsolaritza*:

- Theater plays have predefined scripts, and thus, the improvisation required is very little. The acting person might occasionally change the structure of a sentence but not the meaning. On the other hand, the singed verses must be created in just a moment, according to the requirements imposed by the emcee. There is a strong link to the required form (rhythm, rhymes). As a consequence, a performance is never repeated, it never happens twice the same.
- Plays require dialogues, actors talk to each other or to the public. *Bertsolari-s* mainly sing, but they also need to maintain dialogues with the emcee. Even more, they can interchange messages in form of *bertso-s* with other contestants.
- The scene on the stage changes with the play, but in a verse impromptu performance the verse maker will always find some reference elements like the microphone or the resting chair.
- During a theater play, the public does not participate further than showing the degree of satisfaction with the played stage. On the contrary, in *bertsolaritza* the public can condition the response of the improviser at each moment.

Thus, from the point of view of developing social behaviors in robots, both theater and *bertsolaritza* offer a rich scenario to develop robot expressiveness. The former may require more demanding body language, and the later is better suited to develop human-robot conversation systems. But singing minstrel robots entail social behaviors, robots

must react to perceptions and show to be autonomous. The messages (the verses) need to be created on the spot, based on current perceptions, and the robot needs to adapt to the current situation, to respond to the happening events. But to respond in a natural, human-like manner.

The contribution of this paper relies in designing and implementing the set of behaviors the robots need to show in the stage to increase, in one hand robot autonomy and in the other hand, credibility and sociability.

### III. VERSE GENERATION

When constructing an improvised verse (*bertso*) a number of formal requirements must be taken into account. Rhyme and meter are inseparable elements in improvised verse singing. A person able to construct and sing a *bertso* with the chosen meter and rhyme is considered as having the minimum skills required to be a *bertsolari*. But the true quality of the *bertso* does not only rely on those demanding technical requirements. The real value of the *bertso* resides on its dialectical, rhetorical and poetical value. Thus, a *bertsolari* must be able to express a variety of ideas and thoughts in an original way while dealing with the mentioned technical constraints. In this balance lies the magic of a *bertso*.

#### A. Generating the bertso

*Bertso-s* can be composed in a variety of settings and manners. For instance, *Zortziko Txikia* (see Fig. 2) is a composition of eight lines in which odd lines have seven syllables and even ones have six. The union of each odd line with the next even line, form a strophe. Each strophe has 13 syllables with a caesura after the 7th syllable (7 + 6) and must rhyme with the others. In the basic scenario (the one we'll focus on), the four rhymes to compose a *bertso* are received as input, and the verse generator module should give as output a novel and technically correct verse, and (hopefully) with coherent content. There are other modes but are out of the scope of this paper.

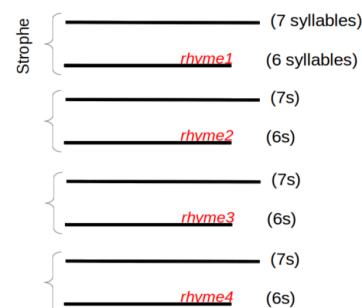


Fig. 2. Structure of a verse in the *Zortziko txikia* meter (8 lines, 4 strophes)

According to Laborde [16], human verse makers have three main tools for improvising verses:

- 1) Learned improvising techniques and rules, mandatory for generating verses metrically correct.

- 2) Memory to store and classify previously listened verses, visual and lexical information.
- 3) The sensorial stimuli that are input in the instants prior to the generation of the verses.

*BertsoBot* has only available the first two tools, the improvisation process is then the result of a set of rules that, given a metric, produce a technically sound verse. And a huge memory, a stored corpus of ordered Basque sentences extracted from a Basque newspaper. Complete sentences need to be stored because they are basic structures that ensure a minimal coherence.

The verse generation process then consists of the following steps:

- 1) Receive as input the four rhymes to compose the verse
  - 2) Find sentences in the corpus that rhyme with the input words and have the correct number of syllables
  - 3) Generate the verse with the highest textual coherence
- See [1] for a more detailed explanation.

### B. Audio processing and singing

In order to generate the verse, the robot needs to identify the proposed exercise and the given rhymes first. The audio is captured via SOX<sup>1</sup> and afterwards, the Google Speech service is used (hopefully available for Basque Language) as ASR to convert the audio to text. Once the text is received, it is analyzed to verify whether the words are available in a local dictionary (list of words with synonyms). If, as a consequence of the analysis no word is recognized, then the robot tells the emcee that it has not understood the sentence and asks to repeat the exercise.

To be able to communicate with the emcee, the robot makes use of AhoTTS, a speech synthesizer for Basque Language developed by AhoLab [11].

But, besides of talking, the robot must sing. The generated verse must be translated to a song in an audio file that will afterwards be reproduced by the robot. To get such audio, first the utilized metric is analyzed and, then, a melody is randomly chosen from an available database and, using a modified version of the AhoTTS that changes the duration and intonation of the syllables, among other features, produces the audio file with the singed verse.

## IV. FIRST PUBLIC PERFORMANCE

At the very beginning of this project we were invited to make a public demonstration: a duel between robots and human *bertsolari*-s. It was a big challenge at the state of the art, and it was an invaluable opportunity not only to make a didactic demonstration of what a real robot could do in *bertso* composition, but also to see how the real *bertsolari*-s, and the illustrated audience will behave and react when faced with synthetic replicates. Fig. 3 shows a snapshot of the event.

The performance aroused great interest, and almost every local newspaper, radio and television covered the event (see

<sup>1</sup>Sound eXchange, a cross-platform command line utility to process audio files



Fig. 3. Verse-duel between one *bertsolari* and two robots

[23], [24]). However, that first performance was a little bit daring, the system development was naive. The verses were improvised, with more or less meaning depending on luck, but the rest of the show, i.e. the robot movements and actions were mostly preprogrammed or teleoperated with a joystick.

Several lessons were extracted from that event. The employed robots, a Pioneer 3DX and a PeopleBot both from MobileRobots, were not very suitable for body language, due to their limited expressiveness. The PTZ unit was used mainly for simulating changes in gaze direction, and small oscillations were implemented to emulate dancing movements while singing.

Beyond the robot morphology, that first performance showed us that a bunch of work was needed before confronting again with human *bertsolari*-s. On the one hand, regarding the verse creation, methods for enhancing verse coherence were needed. On the other hand, the autonomy level of the robots in the stage should be increased and, more important, the way the robots behave on the stage should be humanized. If the robots are meant to participate in such contests, they must show a higher level of expression, much more like human actors do. Next sections show the steps forward being made to improve those behavioral aspects.

## V. BODY LANGUAGE DEVELOPMENT

The *BertsoBot* requires certain capabilities to sing improvised verses to the public, dramatizing the eloquence (gesture repertoire) that a human *bertsolari* shows at the stage. Thus, it should be capable of communicating in a natural way with the emcee and the other contestants, but also to identify some key elements on the stage.

The first decision we made was to change the robotic platforms used. Well, the shape might be not so important but a higher number of degrees of freedom clearly helps. Now, NAO humanoid robots from Aldebaran Robotics are being used.

The verbal and gesture communication capabilities with the new platforms were tested in an initiative named *Zientzi-aClub* or Club of Sciences that aims to disclose science and technologies to the society. A dialogue with NAO of approx. 10 minutes was presented (Fig. 4). The robot was required to give some explanations about itself, and to produce a verse given the rhymes. The robot was teleoperated by human gestures captured by a Kinect sensor (see [27]); NAO

gesticulated while chatting, and moved around the stage according to the teleoperator commands (video available at [25]).



Fig. 4. Dialogue at ZientziaClub. The teleoperator is placed on the same stage, visible to the public

Next subsections explain the modules developed to remove the teleoperation to improve robot autonomy and to supply it with a decent expressiveness. The underlying software architecture is depicted in Fig. 5.

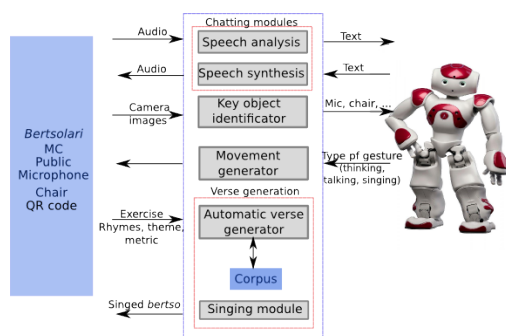


Fig. 5. Software architecture. The modules work in parallel and are activated by different stimuli.

### A. Behavior repertoire

Based on the usual flow of a contest, the robot should be able to:

- 1) Await its turn to sing, until the emcee calls it.
- 2) Approach the microphone and listen to the exercise being proposed to it by the emcee.
- 3) Generate the verse and sing it.
- 4) Observe the public reaction that will allow to feed future verses
- 5) Reach back its chair, or attend to the next exercise according to the emcee's decision

The robot pays attention to different elements at different states. The mic location is a reference point for the robot, and also is the chair. For the time being, those elements, as well as being adapted to the robot morphology, they have labels to make it easier the identification processes. They all have color tags that make them distinguishable; chairs have been painted with different colors and, similarly, the microphone has a blue tag on its base. No location information in form of odometry or frame of reference is used because the location of those elements with respect to the robots varies depending on the scenario. Thus, a color tracking procedure

enhanced with a Kalman Filter is used to produce a more robust behavior against illumination conditions and the robot balancing while walking.

For the microphone tracking, both cameras on the robot head are used. The top camera is used to locate the mic and approach to it. Once the lower camera reaches the view of the microphone, the robot stops forwarding and uses its visual information to correct its position with respect to the microphone.

Besides, for the chair tracking, only the top camera is used to approach it until breast sonars detection alerts that the chair is close enough. Then, the robot turns and uses a yellow line on the floor to center its position with respect to the chair so that it can execute the sitting exercise.

Although most of the time the *bertsolari*-s act individually, sometimes they need to react to other contestant actions. For instance, after one contestant is sent to its chair, and the next one is called, they cannot collide on their trajectories. Humans will naturally do it waiting for the robot or human or letting them pass. But as the system must contemplate the situation with more than a single *bertsolari* robot in a show, the robots need to coordinate among them. At the current state, this coordination is hard coded, there are prefixed timing values set that allow robots to act without pouring into trajectories.

### B. Gesture repertoire

Five different gesture sets have been identified and implemented, using *Choregraphe*<sup>2</sup>, or by modifying some of the movements available within the robot's libraries:

- Thinking gestures: those gestures that, unconsciously, we make while standing up in front of the microphone and thinking the verse. They are movements to unstress, to relax tension like put one's hands back, swing the hip, scratch one's head, ... There is one gesture extremely important while thinking: reach and maintain a neutral pose. The robot needs to move, needs to reproduce some gestures but it cannot be continuously gesturing like a puppet; improvising a verse is a very hard mental process that requires extreme concentration and that is reflected in the body language of the improvisers.
- Talking gestures: humans don't stay still while talking, we naturally gesticulate moving the hands or nodding.
- Singing preamble gestures: just after the improvisation process finishes and before the *bertsolari* starts singing, he/she needs to accommodate the body and/or clear the throat, look around and probably stare off into space, above the public.
- Singing pose: oddly, and probably due to the extreme concentration effort that must be maintained, the *bertsolari* stands still while singing. Of course, not everyone maintains the same pose, sometimes they keep the hands on their pockets, or on their back, or just have their arms down, but that pose does not vary significantly from one *bertsolari* to the other.

<sup>2</sup>A multi-platform desktop application created by Aldebaran for monitoring and controlling NAO humanoid robots



- **Sitting gestures:** humans are not designed to be motionless while being awake, and so, it is not appropriate to have a robot sat inert in the stage. Humans stretch or cross their legs, drink water or move the head to change the gaze while being sat. No need to said that our robots' movements are very limited in that position, and that most of the mentioned moves cannot be replicated. But they can change their arms' position and make movements with their heads. Again, the neutral pose is often required to be maintained.

As Guy Hoffman underlines in [12], when you want to arose emotions, it does not matter so much how something looks like, it is all in the motion, in the timing of how the thing moves. If public attention and interest are to be maintained, gestures cannot be predictable. Even for the robotic enthusiasts, it becomes extremely boring to see the robot doing exactly the same thing once and again. Thus, after identifying the main different states of the global behavior and generating the gesture libraries for each state, we chose to randomly select all, the number of gestures (between a delimited interval), the gesture set and the order in which they must be reproduced, at each state as the performance progresses. The neutral positions while thinking and being sat have a higher probability to be selected due to their importance, and the time to maintain that pose also varies randomly (again within a hard coded time interval).

Regarding the talking and singing states, the duration of the audio file can be measured in advance. Hence, the duration of the associated movement set is adapted to the duration of the audio file.

This solution may seem a little naive, but it has shown to be effective to increase the spontaneity of the robot, from the perspective of the observer and thus, the empathy with the robot.

## VI. DEMOS

We have not had the opportunity to make a public demonstration with the evolved system in a real scenario yet, but the performance of the system can be appreciated in several videos that can be found on our YouTube channel [26]:

a) *Gesture repertoire:* this video reflects different scenes of a play. On the one hand, sitting gestures are demonstrated by two robots that remain sat while gesticulating with different timings and in a different manner. On the other hand, thinking gestures show how the robot behaves while thinking the verse, while talking. Lastly, the singing preamble gestures somehow warn the public it is going to sing.

b) *Behavior repertoire:* this video shows how the robot moves around the stage, when the MC calls it or sends it back to rest.

c) *Chatting and singing behaviors:* the video shows the kind of dialog the robot maintains with the MC in different cases, for instance when it has not been able to understand what the MC has said or how it asks the MC for the rhymes again when it misunderstands them. Besides, the video shows the robot humming when it is not able to compose a strophe with a given rhyme.

d) *Global behavior:* in a rehearsal recorded at the lab, two NAO act as *troubadors* and the roll of the MC is performed by a third robot, a Pioneer 3DX. The robotic MC then establishes the rules of the duel: who starts, the exercises and the flux of the performance. QR codes are used by the emcee to distinguish the two NAO robots (Fig. 6). Verse-maker robots communicate among them sharing messages and each one acts when demanded.



Fig. 6. BertsoBot demo

## VII. SHORTAGES AND FURTHER WORK

It is not easy to objectively evaluate the performance of the proposed system. However, the ontology of robot theater proposed by Lu [19] shall be used to measure the state of the BertsoBot. Lu's ontology is based on the automation level and the required control the robots depend on (see figure 7).

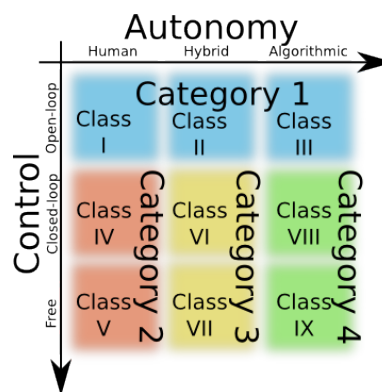


Fig. 7. Ontology of robot theater proposed by David Lu

Analyzing the evolution of the BertsoBot, the first prototype utilized in our first performance could be categorized as a Category 1 Class II robot, an open-loop with a hybrid control, hybrid in the sense that behavior was partially specified by the human, but there were also algorithmically specified behaviors. The second approach, settled with the new platforms and the gesture-based teleoperation could be classified as Category 2 Class IV, a closed-loop system with human input where the performance changes according to some conditions on the stage but not arbitrarily.

The current state of the project locates the BertsoBot at Class VIII, behavior produced algorithmically in a closed-

loop control. The robot generates its behavior via computation, without explicit human intervention further than the oral instructions given by the (robot or human) emcee. But the behavior depends on its own perceptions.

Regarding the behavior of the *BertsoBot* as a single unit, besides improving the verse coherence (some steps forward have been made in [2]), many aspects need to be developed and integrated:

- No many robots show self-awareness of the mistakes they have done (see [28]) and the *BertsoBot* is not an exception. Up to now, if for any reason the robot cannot generate a strophe, the robot hums during that piece of the verse. But the failure is not reflected on the behavior nor in the body expression. We must give the human the sense that the robot knows what it is doing reflecting the errors or the poor actuation sensation on its behavior.
- If the *BertsoBot* is to be trustworthy, public reaction must feedback the robot somehow. For example, in [15] the public is invited to participate showing colored paddles that hints the robot with the kind of jokes the audience (dis)likes.

There is another aspect that affects the coordination of several *BertsoBot*-s acting together that should be improved. Regularly, when it is a human actor that interacts with a robot, she/he tends to adapt to robot timing, filling pauses with her utterances, and helping to conceal delays and robot limitations. But for instance when it is a robot the one that acts as the emcee, the delays (produced by the internet access and the computational units used, but also because of the hard coded timings fixed on the programs) remain. Communication among robots must be extended and more basic behaviors must be integrated.

Summing up, we're still far from having autonomous free robotic *bertsolari*-s (Class IX in Lu's ontology) but we are little by little making steps forward.

#### APPENDIX: COMMUNICATION AMONG ROBOTS

The *BertsoBot* project is being fully developed using ROS ([www.ros.org](http://www.ros.org)), that offers a modular structure. Related packages are available at RSAIT's GitHub ([github.com/rsait/rsait\\_public\\_packages](https://github.com/rsait/rsait_public_packages)). Including a third robot as the MC required to distribute the computation processes and thus, the communication among them. To solve that issue we chose to use `multimaster_fkie`, available in the ROS wiki that allows stabilizing the communication among two or more machines that are running their own roscore.

#### ACKNOWLEDGMENT

This work was supported by the Basque Government Research Team Grant (IT313-10), SAIOTEK Project SA- 2013/00334 and the University of the Basque Country UPV/EHU (Grant UFI11/45 (BAILab))

#### REFERENCES

[1] A. Astigarraga, M. Agirrezabal, E. Lazkano, E. Jauregi, and B. Sierra. BertsoBot: the first minstrel robot. In *6th International Conference on Human System Interaction*, pages 129–136, Sopot, Poland, June 2013. IEEE.

[2] A. Astigarraga, E. Jauregi, E. Lazkano, and M. Agirrezabal. Textual coherence in a verse-maker robot. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*, pages 275–287. Springer International Publishing, Cham, Switzerland, 2014.

[3] C. Breazeal. *Designing sociable robots*. Intelligent Robotics and Autonomous Agents. MIT Press, Cambridge MA, USA, 2004.

[4] K. Čapek. *R.U.R (Rossum's Universal Robots). 1920*. Trans. Paul Selver and Nigel Playfair. Dover Pub. Inc., Mineola, NY, 2001.

[5] J. Fernandez and A. Bonarini. Theatrebot: A software architecture for a theatrical robot. In *Towards Autonomous Robotic Systems*, pages 446–457. Springer, Birmingham, UK, 2014.

[6] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.

[7] J. Garzia. *Jendeaurrean hizlari*. Alberdania, 2008.

[8] P. Gervás. Computational modelling of poetry generation. In *Artificial Intelligence and Poetry Symposium, AISB Convention*, 2013.

[9] M. A. Goodrich and A. C. Schultz. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 1(3):203–275, 2007.

[10] D. Hanson. Hanson Robotics. <http://www.hansonrobotics.com>, [accessed July 20, 2015].

[11] I. Hernaez, E. Navas, J. Murugarren, and B. Etxebarria. Description of the ahotts system for the basque language. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, Perthshire, Scotland, 2001.

[12] G. Hoffman. Robots with "soul". TED talks. January 2014.

[13] B. Li. Robot. ATR. <http://www.geminoid.jp/en/index.html>, [accessed July 13, 2015].

[14] H. Knight. Eight lessons learned about non-verbal interactions through robot theater. *Social Robotics*, pages 42–51, 2011.

[15] H. Knight, S. Satkin, V. Ramakrishna, and S. Divvala. A savvy robot standup comic: Online learning through audience tracking. In *International Conference on Tangible, Embedded and Embodied Interaction*, Funchal, Portugal, 2011.

[16] D. Laborde. *La Mémoire et l'instant. Les improvisations chantées du bertsolari basque*. Elkar argitaletxea, Arsenal Plaza, Baiona, 2005.

[17] B. Li. Robot. <http://www.blancali.com/en/event/99/robot>, [accessed July 21, 2015].

[18] C. Lin, C. Tseng, W. Teng, W. Lee, C. Kuo, H. Gu, K. Chung, and C. Fahn. The realization of robot theater: Humanoid robots and theatrical performance. In *International Conference on Advanced Robotics (ICAR)*, pages 1–6, Munich, Germany, 2009. IEEE.

[19] D. Lu. Ontology of robot theatre. In *Proceedings of the workshop Robotics and Performing Arts: Reciprocal Influences*, ICRA 2012, 2012.

[20] R. Murphy, D. Shell, A. Guerin, B. Duncan, B. Fine, K. Pratt, and T. Zourmos. A midsummer nights dream (with flying robots). *Autonomous Robots*, 30(2):143–156, 2011.

[21] H. G. Oliveira. Automatic generation of poetry: an overview. Technical report, Universidade de Coimbra, 2009.

[22] Pepper. Aldebaran Robotics. <https://www.aldebaran.com/en-a-robots/who-is-pepper>, [accessed July 13, 2015].

[23] RSAIT research group. BertsoBot - galtxagorri eta maialen oinak emanda. <https://www.youtube.com/watch?v=OpQBVmkzRWg>, [accessed July 13, 2015].

[24] RSAIT research group. BertsoBot - tartalo eta egaña puntuka. <https://www.youtube.com/watch?v=x8w4YuNY-Z0>, [accessed July 13, 2015].

[25] RSAIT research group. NAO-scienceclub. <https://www.youtube.com/watch?v=HKxe40-Qi6w>, [accessed July 13, 2015].

[26] RSAIT research group. RSAIT-youtube channel. <https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ>, [accessed July 13, 2015].

[27] I. Rodriguez, A. Astigarraga, E. Jauregi, T. Ruiz, and E. Lazkano. Humanizing nao robot teleoperation using ros. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, pages 179–186. IEEE, 2014.

[28] L. Sharpe. Polite robots show glimmer of self-awareness. <http://www.popsci.com/polite-robots-show-glimmer-self-awareness>, [accessed July 20, 2015].

## 12.2 Minstrel robots: Body language expression through applause evaluation

**Title:** Minstrel robots: Body language expression through applause evaluation

**Authors:** F. Kraemer, I. Rodriguez, O. Parra, T. Ruiz, E. Lazkano

**Conference:** International Conference on Humanoid Robots (Humanoids)

**Publisher:** IEEE

**DOI:** 10.1109/ICRA.2016.7487454

**Year:** 2016

# Minstrel robots: Body language expression through applause evaluation

F. Kraemer<sup>1</sup>, I. Rodriguez<sup>2</sup>, O. Parra<sup>2</sup>, T. Ruiz<sup>3</sup>, E. Lazkano<sup>2</sup>

**Abstract**—Currently humanoid robots become technically more capable of executing complex movements, showing human-like gestures, sometimes even facial expressions, and acting in general. While this lays the basis to make robot theater/enactments more and more interesting for the audience, another key-component is flexibility in the flow of an event to move on from simple pre-scripting. Here a sophisticated method is introduced relying on audio processing, clustering and machine learning techniques to evaluate audience's applauses, allowing the robot to infer self-evaluation about its actions. In a second step we use this information and a humanoid robot's body language to alter the flow of the event and display a reaction for the audience.

## I. INTRODUCTION

Basque, euskara, is the language of the inhabitants of the Basque Country. And *bertsolaritza*, Basque improvised sung poetry, is one of the manifestations of traditional Basque culture that is still very much alive. *Bertso-saio* events, contests in which verse-makers compete, typically feature a number of poets (*bertsolari*-s) that first await a set of words (i.e rhymes), which then should be incorporated in a spontaneously made up poem. The poems are presented to the audience and may make use of one out of various melodies. *Bertsolaritza* offers another sphere to develop robot body language and robot communication capabilities, and thus, to increase robot autonomy and sociability. The *Bertsobot* project [1] aims to develop troubadour robots, and allows for a robot-style enactment. In its current version it features basic poetry creation and follows the different phases of an event, from finding a microphone to presenting a poem. Body language is used to make the robot's actions more lively. A more detailed description of the state of the *Bertsobot* system can be found in [13].

The events usually consist of a rather formal flow of poetry recitations, reasonable appreciation from the audience, mostly by clapping and calming down to silence for the next poet's turn. This situation has been modelled well in *Bertsobot*, but so far the robot had no possibility to show empathy to the audience. If the troubadour performance is to be perceived credible, lively and creative, public reaction must be perceived by the robot somehow and its behaviour

must reflect the noticed sensations, either showing a proper body language or, like real troubadours do, integrating them in the next sung verse.

Despite the thrilled state of the audience, the need for concentration of human poets is very much respected by them, and thus, the crowd waits until the actor finishes to show how pleasant the verses have been, usually clapping as well as laughing when they have found it amusing.

There are several poetry disciplines around the world similar to *bertsolaritza*, the Italian bards, Argentine payadors or Catalan glossators to mention some. But the closest example is the American poetry slam [14] in which poets read or recite poems and are usually judged by selected members of the audience or by a panel of judges. The winner is chosen according to the intensity and duration of the audience's applauses.

The goal of this work is to move on to a closed-loop form of the robot performance where the robot perceives audience's feedback measuring the clapping intensity after it has sung, and then reacting through subtle gestures like it would be expected from a human poet.

## II. RELATED WORK

Developing an approach to react to the audience's feedback covers multiple fields, such as applause detection, classification and selection of the robot's appropriate reaction in the context of the performance.

The problem of content-based audio classification and segmentation has been studied intensively outside the field of robotics and some work has specifically focused on applause. Cai et al. [5] have successfully used Mel-Frequency Cepstral Coefficients (MFCC) and a set of low-level features such as sub-band energies to find significant audience reactions including applause and laughter.

Few work has been done when it comes to observing robot induced audience expressions. Knight et al. [9] have developed a stand-up comedian robot that varies joke selection depending on pre-communicated visual feedback and noise level. Another performance robot by Katevas et al. [8] similarly features joke-telling. It incorporates visual emotion recognition and detecting the noise levels to delay the performing of the comedy script. Audience feedback is partly elicited by the robot itself leaving the spectators in a natural comedy setup without human interference.

Several authors have observed the effect of machines on humans. Nass et al. [12] found that adult humans do not credit anthropomorphic characteristics to computers, although people would still accept questions directly pointing at this and interpret machines in a humanised way. How

\*This work has been partially supported by the Basque Government (IT900-16) and the Spanish Ministry of Economy and Competitiveness MINECO (TIN2015-64395-R)

<sup>1</sup>Author is with Department of Computer Science, University of Freiburg, Germany, kraemerf@informatik.uni-freiburg.de. <sup>2</sup>Authors are with Faculty of Informatics, Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), San Sebastian, igor.rodriguez@ehu.eus. <sup>3</sup>Author is with Faculty of Informatics, Computer Architecture and Technology, University of the Basque Country (UPV/EHU), San Sebastian.



humanoid robots' actions can be designed in order to produce well understandable body language and social cues has been investigated in [2], [6] and [11].

### III. PROPOSED APPROACH

The presented work can generally be split up into a straight-forward workflow. In the initial step, audio processing and machine learning techniques prepare the input audio stream by first chunking it, and then classifying each chunk as being applause or not. Next, the incoming stream of classified chunks is segmented into sections of consecutive applauses, leading to a small descriptor for every evaluated applause. Based on all previous applauses of the event, the most recent one can subsequently be classified, and therefore, a corresponding robot gesture is selected. Fig. 1 summarizes those steps.

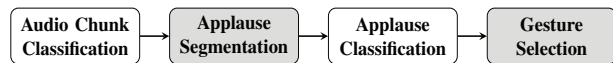


Fig. 1: Approach workflow

Audio chunk classification (section IV) and applause segmentation (section V) are described and evaluated separately. Then, section VI describes how the applause detection is integrated into a live event and the robot behaviour is adjusted to react to the received applauses. It also features elaborated evaluation of the fully developed system in a real event. Finally, we conclude discussing the results and pinpointing future improvements.

### IV. AUDIO CHUNK CLASSIFICATION

Applause detection can be described as a binary classification problem based on the live audio received from the audience. First, a preprocessing step chunks the audio stream into overlapping slices of about 0.1 seconds length using a Hamming-window. Next, the slice is transformed into the frequency domain using a Fast Fourier Transform (FFT) algorithm. And then, the dominant frequency band and Mel Frequency Cepstral Coefficients (MFCC) are extracted using the *Essentia* [4] library.

Some experiments were performed in WEKA [7] to find a suitable approach for the audio classification problem. Several supervised classification algorithms were trained with a database of 5642 audio entries (1169 labelled as applauses and 4473 as non applauses) from heterogeneous *bertso-saio* events: A Support Vector Machine (polynomial kernel, epsilon=1.0E12, complexity 1), Naive Bayes, Bayesian Network (max. 3 parents), 1-Nearest Neighbour and J48 decision tree (conf. factor = 0.25, pruned). 10-fold cross-validation results can be seen in Table I.

	SVM	NB	BN	K-NN	J48
Performance	96.56	95.35	95.53	98.28	97.07
ROC	0.93	0.97	0.97	0.97	0.95

TABLE I: Audio classification comparison

The results show no significant differences among the tested classifiers. K-NN stands out a bit but it is not very well suited for real-time problems. Alternatively, decision trees (J48) are easy to implement and computationally balanced [3]. It must be taken into account that around 20 chunks need to be classified in a second, thus, the classifier needs to give an answer in less than 50 ms. Hence, the J48 decision tree was selected as the final audio classifier. The acoustic energy calculated according to equation 1 and a binary value showing the belonging to the applause class or not for every audio chunk, will be the input for the next step.

$$E = \sum_0^T x(t)^2 \quad (1)$$

$T$  stands for the chunk duration and  $x(t)$  represents the signal value.

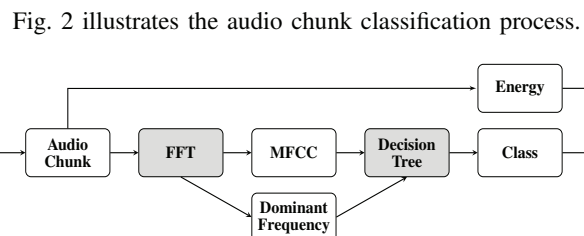


Fig. 2: Audio chunk classification

#### A. Evaluation

The J48 based audio classifier was tested with the following three different collections of audio recordings:

- Cheering and singing from a stadium (labelled as *Stadium*).
- A combination with equally numbered samples from unheard human *bertso-saio* events, event cheers, music and whistle cheers (labelled as *Pos+Neg*).
- A combination of applauses from unheard human *bertso-saio* events superposed with stadium noise (labelled as *BertsoStadium*).

Table II summarizes the results obtained for these test cases.

	Stadium	Pos+Neg		BertsoStadium	
Class	NO	NO	YES	NO	YES
Precision	1.0	0.85	0.76	0.27	0.81
Recall	0.63	0.71	0.85	0.35	0.76
F-Meas.	0.77	0.77	0.80	0.31	0.79

TABLE II: Audio classifier results for test cases

The cheering sounds from the *Stadium* dataset do not contain any applause and are classified with a sufficiently high recall rate. This can be seen as a hard test case since these sounds are unlikely to be heard in a *bertso-saio* event. The *Pos+Neg* dataset shows the quality of the classifier in the event context. The good precision rate proves that it is highly applicable to the problem even for unseen data. Finally, the

superposed audio in *BertsoStadium* shows the limit of the classifier with poor NO-class detection for highly noisy data.

## V. APPLAUSE SEGMENTATION

Once the classification of the chunks has been computed, the next step is to segment the stream of classified chunks and find the portions of applause.

Over the stream of positively or negatively classified audio chunks a sliding window is applied. If the number of positive applause classifications exceeds a certain threshold, the first positive chunk’s start time marks the start of an applause. In case the previous segment was also an applause, a continuation of the applause has been detected. The applause is identified to go on until the percentage of positive applause chunks falls below the threshold. Then, the last positive chunk of the segment marks the end time of the applause. In few occasions applauses can nearly die away and flare up again on the initiative of few individuals. In those cases segments of applauses with little temporal distance (e.g. smaller than 0.5 s) are merged into one. Finally, the energy of each segment is accumulated leading to a 2-dimensional descriptor consisting of an applause’s duration and acoustic energy.

At this point more complex descriptors would be viable, e.g. incorporating more information about applause dynamics like the time-energy relation or dominant frequency bands. Generally, the basic descriptor proved to be sufficient as we are dealing with a rather homogeneous type of applauses. It can then be used to imitate the evaluation strategy commonly used in poetry slams, allowing to judge the performance with an “applause norm”.

### A. Evaluation

To judge the applause segmentation implementation,  $N = 20$  applauses were taken from *bertso-saio* event videos and the detected start and end times were compared to the manually distinguished ones. That is, we considered the observed differences between the true and the program-detected times for applause starting ( $D_{S,i}$ ) as well as for applause ending ( $D_{E,i}$ ). For each type of differences ( $D_j$ ,  $j \in \{S, E\}$ ), the mean ( $\overline{D_j}$ ), Mean Squared Error (MSE) and Mean Absolute Deviation (MAD) statistics have been calculated as follows:

$$\overline{D_j} = \frac{1}{N} \sum_{i=1}^N D_{j,i} \quad (2)$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (D_{j,i} - \overline{D_j})^2 \quad (3)$$

$$\text{MAD} = \frac{1}{N} \sum_{i=1}^N |D_{j,i} - \overline{D_j}| \quad (4)$$

Table III shows that both means are close to zero, showing almost no bias to the true values. The end detection errors are more spread as can be seen by its higher MSE/MAD. This is due to fading out applauses with no hard end bounds.

As a better intuitive measure also the MAD was calculated, which shows that the mean detection error can be expected to be 0.2s to 0.3s for start and end bounds. While these deviation errors can practically add up, they still range low enough for the purposes of this approach, when compared to applause durations usually ranging between 4s and 10s.

	Start Detection Error (s)	End Detection Error (s)
Mean	-0.09	0.03
MSE	0.04	0.22
MAD	0.18	0.33

TABLE III: Applause segmentation errors

## VI. LIVE EVENT

The goal of this work is to make the robot behave accordingly to the audience’s applauses. This is a rather subjective task, first because the high variety of applauses is perceived differently even by humans. And secondly, appropriate reaction gestures must be defined, which need to be well understood by a broad audience.

Subsections VI-A, VI-B and VI-C describe our way to first introduce an objective classification method, and then, choose and execute a suited reaction. Finally, different experiments were performed in real-time in a live event to evaluate the overall robotic system and audience’s acceptance of the robotic performance. In addition, a video of some verses sung by the robot and its reaction to audience’s applauses can be seen at RSAIT’s YouTube channel<sup>1</sup>.

### A. Applause Classification

The applauses are coarsely categorized as belonging to one of the following classes:

- NEGATIVE
- NEUTRAL
- POSITIVE
- VERY\_POSITIVE

The NEGATIVE class is because of the social obligation of also applauding even for poor performance. These “applauses due to politeness” would actually imply a negative feedback for the robot. On the other extreme, very extensive applauses, as they occur at least at the end of an event, also call for an extra class, the VERY\_POSITIVE one.

The applause segmentation step gave us a 2D descriptor containing the duration and energy of the applause. Using this descriptor, now the applause must be classified as belonging to one of the mentioned classes. In general the classification must be done in an online learning fashion. This holds to a greater extent due to the enormous variety in audience sizes, audience knowledge of the robot system, acoustic perception and emotional state of the audience, making distinct events difficult to be compared. Thus, unsupervised online learning techniques fit better to the given problem.

<sup>1</sup>RSAIT’s YouTube channel. <https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ>

We chose to use the k-means [10] algorithm in a non-standard way, as it is one of the simplest algorithms which uses unsupervised learning methods to solve known clustering issues. Instead of fixing the  $k$  value beforehand, and training the algorithm with a database, the algorithm starts with an empty database and new data is incorporated to the training set after each applause session. Subsequently, as new applauses are being perceived, the clustering is executed again, and steadily, the number of available classes is adjusted. The first applause will always be evaluated NEUTRAL. We consider that this first applause means a welcome to the actors and that the show needs a warm up before showing emotional feedback. For the second one two classes are allowed, NEUTRAL or POSITIVE, as audience and robot are getting to know each other and this is the most uncertain learning phase. Afterwards, NEGATIVE, NEUTRAL and POSITIVE classes are offered, until after six feedback rounds sufficient knowledge has been accumulated to also make use of VERY\_POSITIVE class.

As a preprocessing step to k-means both data dimensions are being normalised first, which might be handled differently depending on the event.

Fig. 3 shows an example of the classification of 41 consecutive applauses analysed from a real *bertso-saio* event. There is not a clear separation among classes during the initial steps of the algorithm due to the small amount of data the algorithm is fed with. This is reflected in the online classification results (3a). While the event progresses more data is available and the k-means is able to separate clusters belonging to the four classes (3b).

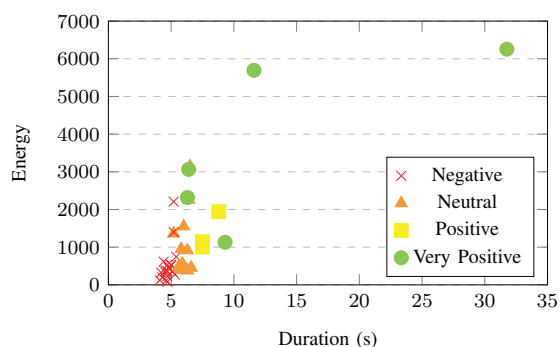
### B. Gesture Selection

For every feedback class a set of 3 predefined gestures has been prepared, giving a total amount of 12 different gestures. Each gesture consists of several movements that must be show some fluency. After the classification of a feedback event, one gesture is randomly chosen out of the corresponding set. To avoid obvious unauthentic behaviour the last executed gesture of the selected class is excluded for the next round, so that there will never be repetitions within a short time period.

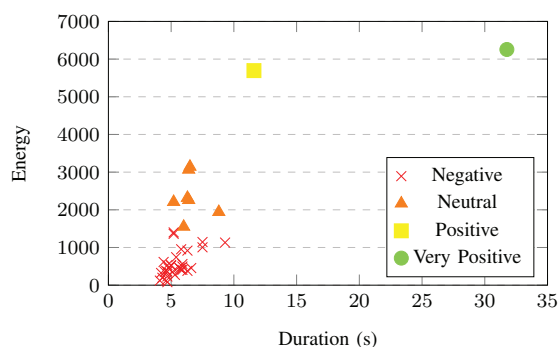
Examples are shown in Fig. 4. The first row corresponds to a negative feedback gesture in which a sad emotion can be clearly appreciated. The sequence relies on a slightly buckled bearing, a shaking of the head and the eyes fixed on the floor. The second row shows a neutral reaction; resting on the body's left side and moving a little its right arm the robot shows indifference. The third row belongs to a positive feedback; a happy reaction can be observed in which the robot moves its hand from bottom to top as celebrating its success. The classic bow shown in the last sequence reflects one of the available very positive reactions.

All gestures usually take 3s to 5s of time and are optionally accompanied with short sounds or Basque phrases. These can range from a reserved "OK, thank you" to a cheerful "Thank you".

Two different behaviours were implemented:



(a) Online learning results



(b) Clustering final state after the last applause

Fig. 3: Applause classification example

- 1) The **diffident behaviour** only makes use of the first three classes, leaving out the VERY\_POSITIVE reaction.
- 2) The **more exaggerated behaviour** aims to clarify the robot's intentions, while it might neglect the typical flow of the event. It makes use of all classes, including the most expressive gestures. Additionally, the robot puts its hand close to its ear claiming for more as soon as it perceives applause sounds.

The first one shows a more restrained or cautious character, while the second one could be categorized as haughty.

### C. Evaluation

Several experiments were conducted in an event to evaluate the overall robotic system and human acceptance of the robotic stage. The audio classification and applause segmentation could be effectively proven to work well with objective offline test input, due to their subjective characteristics. However, the applause classification and gesture selection required to be evaluated online in a live event.

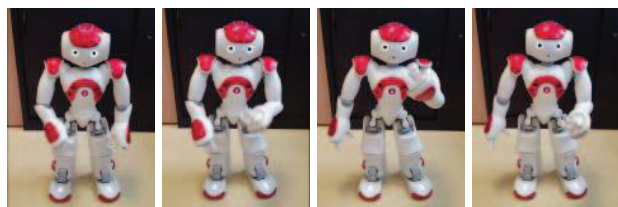
This event was arranged similar to a human *bertso-saio*, with the *bertsolari* robot in front of a seated audience. 17 participants (59% female) familiar with this type of events, took part in listening to the verses and reacting with applauses. After each of the robot's counter-reactions, all participants, lecturers and researchers aged between 25-55, answered a set of questions (see Fig. 5).



(a) Negative



(b) Neutral



(c) Positive



(d) Very Positive

Fig. 4: Examples of robot reaction gestures

The question set was designed carefully to avoid suggestive questions and to get the most honest opinions grading from 1 to 7, 7 being the most positive.

Altogether, 12 *bertso*-s were presented, which were selected from different sources. Four verses were automatically generated by the “automatic *bertso* composer” system [13], which in its current alpha state produces technically permissible but sometimes meaningless poems. Another four were created by non-professional *bertsolari*-s. And the rest were verses composed and sung by professional *bertsolari*-s on national contests. The set of verses was split up into two subsets of 6 and presented to the audience. For the first one only the more diffident behaviour was enacted, while during the second subset the more exaggerated behaviour was used.

Analysing the questionnaires we could infer a lot about the audience’s acceptance of the robotic system during the show. Fig. 6 compares the averages of the different perceptions about the event: how the individuals rated each verse, how they rated the group response and how suitable the gesture

Part 1: Questions to be answered during the show (rate from 1 to 7)

- **Individual:** How much did you like this *bertso*?
- **Group:** According to the applause, how much did the group like it?
- **NAO:** How well did NAO’s reaction fit to the feedback applause?

Part 2: Questions to be answered after the show (rate from 1 to 7)

- How well did NAO understand **positive feedback**?
- How well did NAO understand **negative feedback**?
- Was there **fluency** of motion in the reaction?
- How **expressive** were the reactions?
- How would you rate **variety** of gestures?
- How **close** is the robot behaviour to real *bertso-saio*?

Fig. 5: Summary of the questionnaire

shown by the robot was (Part 1). Text labels correspond to the class of gesture made by the robot.

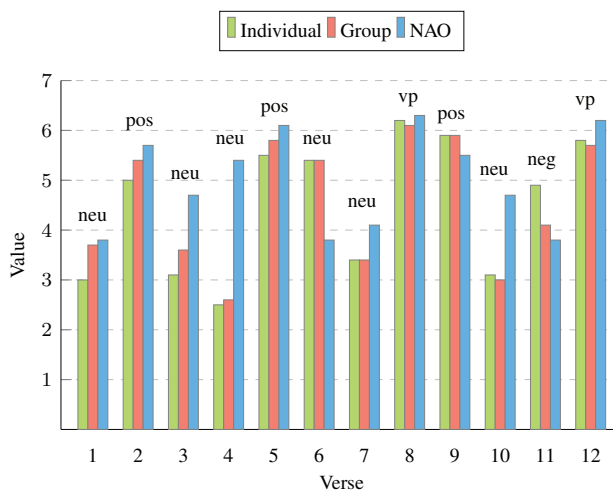


Fig. 6: Audience evaluation of the gesture selection system (answers to part 1 of Fig. 5)

POSITIVE and VERY\_POSITIVE reactions are accepted exceptionally well by the audience. The exception is verse number 6, which was perceived better by the audience, but only classified NEUTRAL. This is because the verse it refers to was enacted at the sixth position, when only three classes were allowed. The only verse (number 11) which was followed by a NEGATIVE reaction, but evaluated quite well by the audience can be explained by problems with the system platform, which led to bad detection of this applause in the first place.

After each subset of verses, some concluding questions about the robot’s behaviour were asked (Part 2). The comparison between the two different behaviours (Table IV) led to the following remarks: While both negative and positive



audience feedback were said to be well understood by the robot, the positive feedback was increased from 5.2 to 6.1 out of 7 possible points with the exaggerated behaviour. The second behaviour also increased the general expressiveness score from 4.8 to 5.9 and the variety improved almost 0.5 to 4.9 points.

When asked about the closeness to a real *bertso-saio* event, the audience rated the first set of *NAO*'s reactions higher (3.9 vs. 3.3), but both scores are below-average.

Question	Diffident beh.	Exaggerated beh.
Positive feedback	5.19 ± 0.83	6.13 ± 0.72
Negative feedback	5.31 ± 1.40	5.14 ± 1.03
Fluency	4.75 ± 1.13	4.71 ± 0.99
Expressiveness	4.75 ± 1.29	5.87 ± 0.92
Variety	4.44 ± 1.03	4.88 ± 1.45
Closeness	3.94 ± 1.12	3.31 ± 1.45

TABLE IV: Public response average values for the Part2 of the questionnaire

In order to get more insight about the quality of the gesture set, the audience was confronted to each gesture in a random sequence while asked for a gesture tag. Fig. 7 shows the measured classification rates. Only case 5 was wrongly classified as being NEUTRAL for the big majority of the audience, while the gesture was POSITIVE. Cases 2, 3, 6, 8 and 12 do not show a clear definition, as the results show a tie between the right tag and a neighbour class.

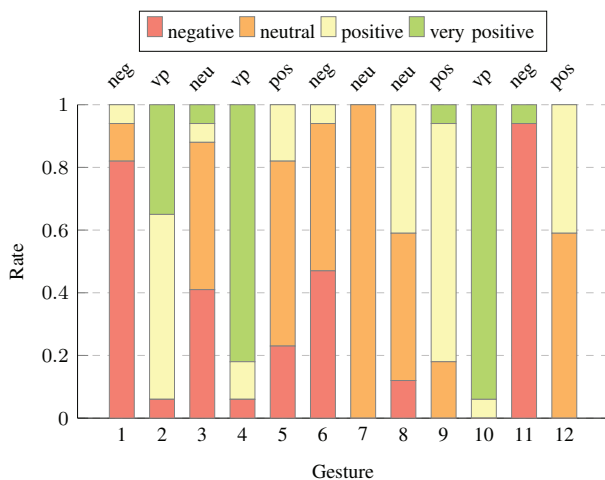


Fig. 7: Audience gesture classification rates for the 12 gestures available. The execution order was randomly selected

## VII. CONCLUSION AND OUTLOOK

Without questioning for additional information or unusual behaviour of the audience, this work already allows robots to react and alter their behaviour during an event according to a specific audience's natural feedback. We have found that the audience felt better understood when the robot exaggerated its behaviour. It is inconceivable for a *bertsolari* to show any kind of arrogance in such a traditional cultural event that requires extreme concentration. As a consequence, this

result makes us think that we should detach the robot event from its human counterpart. Moreover, it appears that instead of imitating the original event in detail, a new identity should be established for the robot.

The classification of the applause could also be achieved by a Gaussian Mixture Model allowing for a more sophisticated classification using priors. A comparison with the already implemented system should be carried out. Another extension may be to deduce the number of needed gesture classes from an error measure over the current classification. Future work will also include investigation about audience-sensitive planning and integration of further, but more subtle human feedback, like emotions in general through facial expressions or reservation through delayed reactions. Also we will research more in the field of gesture selection evaluating the possibility of movement styles combined with probabilistic methods. The robot's empathetic level would be improved by applying sentiment analysis to the verses sung by the opponents and combining the results with the proposed gesture selection mechanism.

## REFERENCES

- [1] A. Astigarraga, M. Agirrezabal, E. Lazkano, E. Jauregi, and B. Sierra. Bertsobot: the first minstrel robot. In *Human System Interaction (HSI), 2013 The 6th International Conference on*, pages 129–136. IEEE, 2013.
- [2] A. Beck, B. Stevens, K. A. Bard, and L. Cañamero. Emotional body language displayed by artificial agents. *ACM Transactions on Interactive Intelligent Systems (TüS)*, 2(1):2, 2012.
- [3] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, New York, 2007.
- [4] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra. Essentia: an audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR'13)*, pages 493–498, Curitiba, Brazil, 04/11/2013 2013.
- [5] R. Cai, L. Lu, H. Zhang, and L. Cai. Highlight sound effects detection in audio stream. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 3, pages III–37. IEEE, 2003.
- [6] S. Embgen, M. Luber, C. Becker-Asano, M. Ragni, V. Evers, and K. O. Arras. Robot-specific social cues in emotional body language. In *RO-MAN, 2012 IEEE*, pages 1019–1025. IEEE, 2012.
- [7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.
- [8] K. Katevas, P. G. T. Healey, and M. T. Harris. Robot stand-up: Engineering a comic performance. In *2014 Workshop on Humanoid Robots and Creativity*, 2014.
- [9] H. Knight, S. Satkin, V. Ramakrishna, and S. Divvala. A savvy robot standup comic: Online learning through audience tracking. *Workshop paper (TEI'10)*, 2011.
- [10] J. N. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.
- [11] D. McColl and G. Nejat. Recognizing emotional body language displayed by a human-like social robot. *International Journal of Social Robotics*, 6(2):261–280, 2014.
- [12] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.
- [13] I. Rodríguez, A. Astigarraga, T. Ruiz, and E. Lazkano. Singing minstrel robots, a means for improving social behaviors. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 2902–2907. IEEE, 2016.
- [14] S. B. A. Somers-Willett. *The cultural politics of slam poetry: Race, identity, and the performance of popular verse in America*. University of Michigan Press, 2009.

## 12.3 BertsoBot: Towards a Framework for Socially Interacting Robots

**Title:** BertsoBot: Towards a Framework for Socially Interacting Robots

**Authors:** A. Astigarraga, I. Rodriguez, T. Ruiz, E. Lazkano

**Conference:** Jornadas Nacionales de Robótica

**Year:** 2017

## BertsoBot: Towards a Framework for Socially Interacting Robots

A. Astigarraga<sup>a,1,2\*</sup>, I. Rodríguez<sup>a</sup>, T. Ruiz<sup>b</sup>, E. Lazkano<sup>a</sup>

<sup>a</sup> Computer Sciences and Artificial Intelligence, University of Basque Country (UPV/EHU), Manuel Lardizabal 1, 20018 Donostia.

<sup>b</sup> Computer Architecture and Technology, University of Basque Country (UPV/EHU), Manuel Lardizabal 1, 20018 Donostia.

### Abstract

The objective of this article is to compile the work carried out by the RSAIT<sup>3</sup> research group on the BertsoBot project. The BertsoBot project aims to develop an autonomous robot capable of composing and playing traditional Basque impromptu verses – bertsoak. The developed system is able to construct novel verses according to given constraints on rhyme and metric that also show semantical coherence, and to perform it in public. The BertsoBot project, at the intersection of Autonomous Robotics, Natural Language Generation and Human Robot Interaction, works to model the human abilities that collaborate in the process of creating and performing impromptu verses in front of an audience. This paper brings together the steps taken in the design and implementation of robot's individual behaviors and the overall control architecture. Copyright © 2017 CEA.

### Palabras Clave:

Robotics, Natural Language Processing, Machine Learning.

### Datos del Proyecto:

Denominación del proyecto: Análisis de Personas con Biometría Blanda para Servicios Inteligentes Multilingües

Referencia: TIN-2015-64395-R

Investigador/es responsable/es: Basilio Sierra y Elena Lazkano

Tipo de proyecto (internacional, nacional, autonómico, transferencia): Nacional

Entidad/es financiadora/s: Ministerio de Economía y Competitividad

Fecha de inicio/fin: 01/01/2016 – 31/12/2018

### 1. Introduction

Until recently, it was impossible to consider humans and robots living together. But now, robots start to become companions or co-workers of humans, opening an important research domain to build robots that are able to intuitively interact with humans. A considerable number of robotic systems have been developed in the last decade showing Human Robot Interaction (HRI) capabilities [Fong et al., 2003][Goodrich and Schultz, 2007].

However, social robots are beyond HRI. According to Breazeal [Breazeal, 2004], sociable robots are socially intelligent robots in a human like way, and they need to show the “human social” characteristics like the expression of emotions, the ability to conduct high-level dialogue, to learn, to develop personality, and to develop social competencies. In consonance with FeilSeifer and Mataric [Feil-Seifer and Mataric, 2011] social robots can be categorized as assistive robots (AR), socially interactive robots (SIR) and socially assistive robots (SAR). Regardless of the applications, in the

last years, research in the field of social robotics has grown. Several robots have been designed in this area, to support development of self-efficacy and emotional well-being in diabetic children [Cañamero and Lewis, 2016], as interactive teachers in a collaborative learning class with infants [Kanda et al., 2012] and as shopping mall guide, designed for customer navigation, information providing and enjoyment [Chen et al., 2015], to mention some.

Entertainment is an area in which social robots can have high impact. Public performances using robots have shown to be a great setting for disclosing the state of the art of social robots to the general public. Theatre, a live entertainment activity, offers an invaluable field to research and develop social skills in robots. Although everything is rehearsed beforehand, theatre offers an invaluable sphere to research and develop social behaviours in robots, to work and extend the expression of emotions and the natural communication among humans and robots [Lin et al.2009] [Fernandez and Bonarini, 2014]. A review of robot performances can be found in [Murphy et al., 2011]. Little by little robots are bursting into

<sup>1</sup> Aitzol Astigarraga.

<sup>2\*</sup> Autor en correspondencia.

Correos electrónicos: aitzol.astigarraga@ehu.eus (A. Astigarraga),

igor.rodriguez@ehu.eus (I. Rodríguez),

txelo.ruiz@ehu.eus (T. Ruiz),

e.lazkano@ehu.eus (E. Lazkano)

URL: <http://www.sc.ehu.es/ccwrobot/seccion/members-2/subseccion/aitzol-astigarraga-2> (A. Astigarraga)

<sup>3</sup><http://www.sc.ehu.es/ccwrobot/seccion/home/lang/en>

theatres motivated by researchers as a means, but also by artists [Li].

The *BertsoBot* project described in this paper resumes the work done by RSAIT in the last years. This project provides a huge opportunity to develop social robot capabilities in the context of *bertsolaritza*, a traditional Basque improvised singing poetry manifestation. It aims to develop minstrel robots that, beyond generating verses automatically, are able to sense the environment and interact with it, and show proper body language and robot communication skills, like real troubadours do. Aldebaran's NAO humanoid robotic platforms are being used that, although faceless, allow for body language development and have multisensory capabilities.

## 2. Bertsolaritza and Automatic Verse Generation

Basque, *Euskara*, is the language of the inhabitants of the Basque Country. And *bertsolaritza*, Basque improvised contest poetry, is one of the manifestations of traditional Basque culture that is still very much alive.

Events and competitions are very common (see Figure 1), which usually consist of a rather formal flow of poetry recitations, *bertso*-s. In such performances, several verse-makers, compete with each other singing improvised verses about topics or prompts which are given to them by an emcee (theme-prompter). They compose the verses on the fly, normally in less than one minute, and sing a poem along the pattern of a prescribed verse-form that also involves a rhyme scheme. Melodies are chosen from among hundreds of tunes. Xabier Amuriza, a famous verse-maker defined *bertsolaritza* in a verse as:

<p><i>Neurritz eta errimaz kantatzea hitza horra hor zer kirol mota den bertsoari.</i></p>	<p><b>Through meter and rhyme at singing a word could bertsoari be seen as sport.</b></p>
--	---



Figure 1: 2015 National bertsoari's championship.

Different poetry disciplines similar to *bertsolaritza* can be found around the world, such as Catalan glossators, Argentine payadors or Italian bards to mention some. However, the closest example is the American poetry slam (Somers-Willett, 2009) in which poets read or recite poems and are sometimes judged by selected members of the audience and sometimes, like in *bertso* contests, by a panel of judges.

The art of composing improvised verses requires a number of prerequisites that must be taken into account. We can say that any person with the capabilities to construct and sing a *bertso* with the chosen meter and rhyme has the minimum skills to be a *bertsolari*. But the real value of the *bertso* goes beyond composing a verse according to those demanding

technical requirements. Its real value resides on its dialectical, rhetorical and poetical value [Garzia et al. 2001]. Thus, a *bertsolari* must be able to express a variety of ideas and thoughts in an original way while dealing with the mentioned technical constraints. In this balance lies the magic of a *bertso*.

*Bertso*-s can be composed in a variety of settings and manners. For instance, *Zortziko Txikia* (see Figure 2) is a composition of eight lines in which odd lines have seven syllables and even ones have six. The union of each odd line with the next even line form a strophe. Each verse has 13 syllables with a caesura after the 7th syllable (7 + 6) and must rhyme with the others.

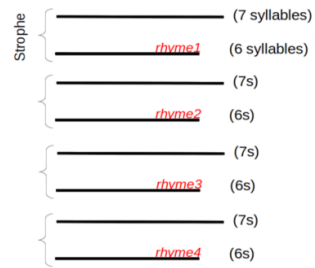


Figure 2: Structure of a verse in the Zortziko txikia meter.

### 2.1. Automatic Verse Generation

Computer-based poetry has been paid attention to in the research community for the last years (see [Gervás, 2013] and [Oliveira, 2009] for a review), but among the several differences that exist between poetry and *bertsolaritza*, mainly the later belongs to the oral genre, and the public performance is extremely important.

According to Laborde [Laborde, 2005], human verse makers have three main tools for improvising verses:

1. Learned techniques and rules for improvisation, mandatory for generating verses metrically correct.
2. Memory to store and classify previously listened verses, visual and lexical information.
3. The sensorial stimuli that are input in the instants prior to the generation of the verses.

The improvisation process is then the result of a set of rules that, given a metric, produce a technically sound verse with content obtained from a huge memory.

We have developed two poem generation strategies that respond to popular exercises in *bertsolaritza*:

- 4 rhymes given: four rhyming words are given and is required to compose the bertso "around" these rhyming words.
- Theme given: a bertso must be composed on the given subject.

In this work we will focus on the first strategy, and thus, the automatic verse generation process consists of the following steps:

1. Receive as input the four rhymes to compose the verse.
2. Find sentences in the corpus that rhyme with the input words and have the correct number of syllables.
3. Generate the verse with the highest textual coherence.





### 4.3. Interaction with the environment

#### 4.3.1. Face and Sound localization

A natural reaction when we want to interact with someone is to direct our gaze towards the interested agent. The gaze feeds the communication, and conveys interest or attention to the interlocutor. It requires positioning the robot to make the most out of its sensors and to let the human talker know what the robot is actually paying attention to. Spontaneity during verbal communication involves two main behaviours, face and sound localization.

Face localization is done applying OpenCV's Haar feature-based cascade classifiers [Viola and Jones, 2001] to the images taken by the upper camera on the NAO's head. Once the face is detected within an image, the center of the face in the image is obtained, and the head joint angles to track the face, with respect to the center of the image are calculated.

Sound localization allows a robot to identify the direction of sound, and it is done using Aldebaran's "ALSoundDetection" algorithm based on TDOA (Time Difference of Arrival) approach [Bensky, 2016]. The sound wave emitted by a source is received at slightly different times on each of the NAOs four microphones, from the closest to the farthest. These differences are related to the current location of the emitting source. By using this relationship, the robot is able to retrieve the direction of the emitting source (azimuth and elevation angles) from the TDOAs measured on the different microphone pairs.

#### 4.3.2. Find key objects

The robot pays attention to different elements at different states. The robot can be requested to reach the microphone to start its singing turn or it may need to go to rest to its chair. For the time being, those elements, as well as being adapted to the robot's morphology, they have labels to make it easier the identification and recognition processes. They all have colour tags that make them distinguishable; chairs have been painted with different colours and, similarly, the microphone has a blue tag on its base. Every key object has a QR code to make it recognizable. A colour tracking procedure enhanced with a Kalman Filter is used to produce a more robust behaviour against illumination conditions and balancing produced during walking. No location information in form of odometry or frame of reference is used because the location of those elements with respect to the robots varies depending on the scenario.

#### 4.3.3. Feedback from the audience

Audience plays an important role in any type of performances, specially in *bertsolaritza*. Despite the thrilled state of the audience, the need for concentration of human poets is very much respected by them, and thus, the crowd waits until the actor finishes to show how pleasant the verses have been, usually clapping as well as laughing when they have found it amusing.

Perceiving and showing emotions is essential to convey interaction. The Affective Loop is the interactive process in which the user of the system first expresses her/his emotion through some physical interaction involving her

body, and the system responds by generating affective expression, which in turn affects the user making her/him respond and step-by-step feel more and more involved with the system [Höök, 2009][Paiva et al., 2015].

Developing an approach to react to the audience's feedback covers multiple fields, such as applause detection, classification and selection of the robots appropriate reaction in the context of the performance.

The presented approach uses audience applause as feedback to the robot system [Kraemer et al., 2016]. Applauses are captured and translated into a response from the public by means of energy (E) and duration (d) of the applause. The addressed strategy can be split up into a straight-forward workflow (see Figure 4). In the initial step, audio processing and machine learning techniques prepare the input audio stream by first chunking it, and then classifying each chunk as being applause or not. Next, the incoming stream of classified chunks is segmented into sections of consecutive applauses, leading to a small descriptor ([E, d]) for every evaluated applause. Based on all previous applauses of the event, the most recent one can subsequently be classified. The applauses are coarsely categorized as belonging to one of the following classes: *Negative*, *Neutral*, *Positive* and *Very Positive*.



Figure 4: Approach workflow

As all events are different, it is difficult to make a comparison between them. Thus, in our approach we use unsupervised online learning techniques. We use k-means to do clustering with a variable number of classes. The first applause is always classified as *Neutral*, after that, the number of available classes is increased.

### 4.4. Motion and Gestures controller

#### 4.4.1. Robot State

"Robot State Controller" provides information about the posture in which the robot is, and whether the robot is moving or not. The posture classification is made by a C4.5 decision tree [21] that classifies robot posture as being sat on floor, sat on chair, crouched (rest position) and standing up. Postures were represented using 150 variables obtained from the TF<sup>7</sup> tree of each robot joint (x, y, z, roll, pitch, yaw) values of the 25 DoF). To train the C4.5 classifier a total of 240 data entries were collect, 60 entries of each class. In order to determine if the robot is moving or not a comparison of all joints is done in two different times.

#### 4.4.2. Motion

There are several behaviours that output a motion action. For instance, when the emcee's face or a sound is detected the robot must direct its gaze and body to her/him, when it is its singing turn it must walk to the microphone, or when receiving

<sup>7</sup><http://wiki.ros.org/tf>

a specific order in the context of the performance. All these motion actions are managed by a robot motion controller which makes the robot react accordingly to the current situation, and which is also conditioned by the robot actual state.

#### 4.4.3. Gestures

If our robots are meant to participate in such verse contests, beyond singing capabilities, they must show the same degree of expressiveness Basque troubadours do. Improvising a verse is a very hard mental process that requires extreme concentration and that is reflected in the body language of the improvisers. When the *bertsolari*-s are on the stage they are continuously conveying information, through facial expressions, body postures, movements or gestures, intentionally or not, about their emotional state.

The robot needs to move, needs to reproduce some gestures but it cannot be continuously gesturing like a puppet. After identifying the main different states of the global behaviour, a gesture library composed by five different gesture sets have been defined to mimic troubadours' emotional behaviour on the stage [Rodriguez et al. 2016]. At each state of the performance appropriate gesture set is selected, and in order to avoid to see the robot doing exactly the same thing once and again, the gesture to be reproduced, the interval between gestures, the execution order and the number of gestures are always randomly selected. The five gesture sets are:

- *Thinking gestures*: Those gestures that, unconsciously, humans make while standing up in front of the microphone and thinking the verse. They are movements to unstress, to relax tension like put one's hands back, swing the hip, scratch one's head, etc. There is one gesture extremely important while thinking: reach and maintain a neutral pose.
- *Talking gestures*: Humans do not stay still while talking, we naturally gesticulate moving the hands or nodding. NAO accompanies its speech moving its arms too. In this case, the number of gestures to perform varies according to the duration of the speech. The robot also nods to make visible that it has understood something. It does not mean that it knows what has been said, but it makes the interlocutor realize that the robot has successfully processed the captured audio.
- *Singing preamble gestures*: Just after the improvisation process finishes and before the *bertsolari* starts singing, he/she needs to accommodate the body and/or clear the throat, look around and probably stare off into space, above the public. Oddly, and probably due to the extreme concentration effort that must be maintained, the troubadours stands still while singing. Of course, not everyone maintains the same pose, sometimes they keep the hands on their pockets, or on their back, or just have their arms down, but that pose does not vary significantly from one *bertsolari* to the other. Thus, no gesture is reproduced while singing.
- *Waiting gestures*: Humans are not designed to be motionless while being awake, and so, it is not appropriate to have a robot sat inert or stand up

paralyzed in the stage. Humans stretch or cross their legs, drink water or move the head to change the gaze while being sat. No need to say that our robots' movements are very limited in that position, and that most of the mentioned moves cannot be replicated. But they can change their arms' position and make movements with their heads. Again, the neutral pose is often required to be maintained.

- *Emotional reaction gestures*: After the *bertsolari* sings a verse the audience responds applauding to express their opinion, and this reaction is reflected in the robot as emotion gesture. Each applause feedback class has been represented with an emotion; in the next order *Sad*, *Calm*, *Joy* and *Excited* emotions correspond to *Negative*, *Neutral*, *Positive* and *Very Positive* applause classes. Two different behaviours were implemented. The diffident behaviour only makes use of the first three classes. The more exaggerated behaviour makes use of all classes.

## 5. Public Performances

The robots' performance capabilities have been demonstrated in different events in a 4 years period. These public performances show the evolution of the *BertsoBot* project since its start up, when no humanoid platform was available and up to now.

- **2012/04 - First public appearance**: Inauguration of the speaker's corner of our Campus. Paradoxically the most audacious one, due to the importance of the event and the preliminary state of the project. Tartalo and Galtxagorri – PeopleBot and Pioneer2DX platforms – were brought out and acted outdoor. No significant body language was shown, neither chatting was possible. Robots were mainly teleoperated and control software was Player/Stage. Only the automatic verse generation system was embedded in wheeled robots. Video available<sup>8</sup>.
- **2013/05 - Robots against bachelor students**: An event hold in the Faculty of Informatics where robots competed against some *bertso*-amateur students. Tartalo was accompanied by NAO for the first time. Primary gestures were shown by NAO, that acted as the emcee semi-autonomously. NAO was controlled using Choregraphe. Video available<sup>9</sup>.
- **2014/03 - Women's day at the Faculty**: The UPV/EHU annually celebrates the women's international day in a different center and in 2014 it was held at our Faculty. The program included a *bertso* event where two big professionals and two robots (NAO and Tartalo) took part. NAO showed improved chatting abilities, but still "unROSified". Primary gestures in NAO were shown, which guided the event but semi-autonomously<sup>10</sup>.

<sup>8</sup><https://www.youtube.com/watch?v=OpQBVmkzRWg>

<sup>9</sup><http://www.eitb.eus/eu/kultura/bertsolaritza/osoa/1350970/robotbertsolariak-ixa-taldea-eta-ehuko-robotika-saila/>

<sup>10</sup><http://ehutb.ehu.es/es/video/index/uuid/531ec65f964be.html>

- **2014/11 - ScienceClub:** Club of Sciences events aim to disclose science and technologies to the society. A dialogue with NAO entitled “Chatting with NAO” of approximately 10 minutes was presented. NAO acted alone and it was its first performance after being “ROSified”. However, it still acted semi-autonomously.
- **2015/11 - ScienceClub:** Next year the title of the event was “NAO, an empathetic or just amusing robot?”. Body gestures were integrated and chatting abilities were shown. The key object recognition was tested together with the face and sound localization behaviours. Video available<sup>11</sup>.
- **2016/02 - Discrete event at the Faculty:** A discrete event was organized at the Faculty in order to be able to evaluate the applause classification and emotional state gesture reproduction modules. Thinking and singing preamble gestures were used. Video available<sup>12</sup>.
- **2016/09 - Closing of a Summer University Course: BertsoBot** was invited to the closing of a course entitled “Educational assessment: unresolved matter” (organized by the University of Basque Country). It was not a *bertso-saio* event but it covered all aspects of the interaction. It was a short exhibition in which NAO sang only one verse and thus, the applause feedback only allowed to reflect a Calm state. Unfortunately, we have no media of the event.
- **Lab demonstration:** A rehearsal without audience recorded at our laboratory<sup>13</sup> exhibits the global behaviour of the *BertsoBot* system in a performance similar to *bertsolaris* events, in which two NAOs act as troubadours and the roll of the emcee is performed by Galtxagorri. The robotic emcee establishes the rules of the duel: who starts, the exercises and the flux of the performance.

## 6. Further Work

The work carried out during this project has revealed many promising areas of further research, such as in computer-based poetry and social robotics fields.

Regarding to the further research in computer-based poetry, we are working to improve the verse generation module by generating impromptu verses using Markov chains, and applying sentiment analysis to build higher quality poems.

On the other hand, we are considering to dynamically adapting the different DoFs to express emotions instead of precompiled gestures.

## Acknowledgements.

This work has been partially supported by the Basque Government (IT900-16), the Spanish Ministry of Economy and Competitiveness MINECO/FEDER (TIN 2015-64395-R). Author Rodriguez has received the UPV/EHU research Grant PIF13/104.

11<https://www.youtube.com/watch?v=IMMXHWB2mZA>

12<https://www.youtube.com/watch?v=SdxNgmV3CzA>

13<https://www.youtube.com/watch?v=UNhvd2qbuaY>

## References

- Astigarraga, A., Agirrezabal, M., Lazkano, E., Jauregi, E., and Sierra, B., 2013. BertsoBot: the first minstrel robot. In 6th International Conference on Human System Interaction, 129–136.
- Astigarraga, A., Jauregi, E., Lazkano, E., and Agirrezabal, M., 2014. Textual coherence in a verse-maker robot. In Human-Computer Systems Interaction: Backgrounds and Applications 3, 275–287.
- Bensky, A., 2016. Wireless positioning technologies and applications. Artech House.
- Breazeal, C., 2004. Designing sociable robots. Intelligent Robotics and Autonomous Agents. MIT Press.
- Cañamero, L. and Lewis, M., 2016. Making “new ai” friends: Designing a social robot for diabetic children from an embodied AI perspective. Social Robotics, 8:523–537.
- Chen, Y., Wu, F., Shuai, W., Wang, N., Chen, R., and Chen, X., 2015. Kejia robot—an attractive shopping mall guider. In International conference on Social Robotics, 145–154. Springer.
- Feil-Seifer, D. and Matari c, M. J., 2011. Socially assistive robotics. IEEE Robotics & Automation Magazine, 18(1):24–31.
- Fernandez, J. and Bonarini, A., 2014. Theatrebot: A software architecture for a theatrical robot. In Towards Autonomous Robotic Systems, 446–457. Springer.
- Fong, T., Nourbakhsh, I., and Dautenhahn, K., 2003. A survey of socially interactive robots. Robotics and autonomous systems, 42(3):143–166.
- Garzia, J., Sarasua, J., & Egaña, A., 2001. *The art of bertsoaritzas: improvised Basque verse singing*. Bertsoari liburua.
- Gervás, P., 2013. Computational modelling of poetry generation. In Artificial Intelligence and Poetry Symposium, AISB Convention.
- Goodrich, M. A. and Schultz, A. C., 2007. Human-robot interaction: a survey. Foundations and trends in human-computer interaction, 1(3):203–275.
- Hernaiz, I., Navas, E., Murugarren, J., and Etxebarria, B., 2001. Description of the ahotts system for the basque language. In 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis.
- Höök, K., 2009. Affective loop experiences: designing for interactional embodiment. Philosophical transactions of the royal society B, 364:3585–3595.
- Kanda, T., Shimada, M., and Koizumi, S., 2012. Children learning with a social robot. In Human-Robot Interaction (HRI), 351–358. IEEE.
- Kraemer, F., Rodriguez, I., Parra, O., Ruiz, T. and Lazkano, E., 2016. Minstrel robots: Body language expression through applause evaluation. In *Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on* (pp. 332-337). IEEE.
- Laborde, D., 2005. La M emoire et l’instant. Les improvisations chant ees du bertsoari basque. Elkar.
- Li, B. Robot. <http://www.blancali.com/en/event/99/robot>.
- Lin, C., Tseng, C., Teng, W., Lee, W., Kuo, C., Gu, H., Chung, K., and Fahn, C., 2009. The realization of robot theater: Humanoid robots and theatric performance. In International Conference on Advanced Robotics (ICAR), 1–6.
- Murphy, R., Shell, D., Guerin, A., Duncan, B., Fine, B., Pratt, K., and Zourntos, T., 2011. A midsummer nights dream (with flying robots). Autonomous Robots, 30(2):143–156.
- Oliveira, H. G., 2009. Automatic generation of poetry: an overview. Technical report, Universidad de Coimbra.
- Paiva, A., Leite, I., and Ribeiro, T., 2015. The Oxford Handbook of Affective Computing, chapter Emotion Modelling for Social Robots. Oxford University Press.
- Quinlan, J. R., 2014. C4. 5: programs for machine learning. Elsevier.
- Rodriguez, I., Astigarraga, A., Ruiz, T., and Lazkano, E., 2016. Singing minstrel robots, a means for improving social behaviors. In *Robotics and Automation (ICRA)*, (pp. 2902-2907).
- Somers-Willett, S. B., 2009. The cultural politics of slam poetry: race, identity, and the performance of popular verse in America. University of Michigan Press.
- Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001, volume 1, 1–511.

## 12.4 Body Self-Awareness for Social Robots

- Title:** Body Self-Awareness for Social Robots
- Authors:** I. Rodriguez, A. Astigarraga, T. Ruiz, E. Lazkano
- Conference:** International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)
- Publisher:** IEEE
- DOI:** 10.1109/ICCAIRO.2017.23
- Year:** 2017

# Body Self-Awareness for Social Robots

Igor Rodriguez\*, Aitzol Astigarraga\*, Txelo Ruiz<sup>†</sup> and Elena Lazkano\*

\*Department of Computer Science and Artificial Intelligence, Faculty of Informatics  
University of Basque Country (UPV/EHU), Donostia  
Email: igor.rodriguez@ehu.eus

<sup>†</sup>Department of Computer Architecture and Technology, Faculty of Informatics  
University of Basque Country (UPV/EHU), Donostia  
Email: txelo.ruiz@ehu.eus

**Abstract**—Just as humans show conscious of their body, social robots, in the way to be truly autonomous, they should also be able to recognize its own configuration. Our research group is working on a project named *BertsoBot* which aims to develop social minstrel robots for entertainment. The work presented here focuses on the automatic recognition of robot self body postures. Only proprioceptive information is being used and several supervised classifiers are compared to make the appropriate choice that fulfills the task requirements. A ROS module that performs the online classification has been implemented for endowing the robot with self awareness capabilities. The developed implementation allows our NAO minstrel robot to make decisions based on its body posture and state instead of just relying on a blind finite state automaton. A demo is provided in a link to a video.

## I. INTRODUCTION

According to Lagercrantz et al. [13] a simple definition of consciousness is the sensory awareness of the body, the self, and the world. The first thing the child perceives is his own body, which serves as a means of interaction with others and the environment. Thanks to her/his body, the child experiences different sensations, mobilizes and learns [11]. Recognizing oneself, although it seems easy, at least requires to be conscious of ones body and ones actions.

The five senses (sight, hearing, smell, touch and taste) are the traditionally recognized methods of perception responsible for our interaction with the external world. Additionally, we have several senses that are responsible for our internal functioning. One of the most important internal senses is called proprioception, which provides feedback about the status of the body internally [12]. Thanks to muscle spindles, which detects changes in the muscle and signal the angle of related joints, we got information about limb positions, and we realize our body's position.

Just as humans show conscious of their body, social robots, in the way to be truly autonomous, should also be able to recognize its configuration. Several robotic systems can be considered as self-aware systems, able to recognize themselves in the mirror [10], able to be aware of their motion [14], or able to change their own models of their physical embodiment [4], to mention some.

We are working on a project named *BertsoBot* which aims to develop social minstrel robots for entertainment [2] [18]. Our NAO humanoid robot, as a minstrel, is able to interact with the public and sing improvised verses showing the expressiveness

of Basque troubadours at the stage to a certain extent. The work presented here focuses on the automatic recognition of robot self body postures. Taking into account only internal sensory receptors the posture recognition system developed endows the robot with the ability to know its body posture and distinguish whether it is moving or not without the aid of visual information. According to the posture detected, the robot knows what type of movements or actions it can perform.

The rest of the paper is structured as follows: Section II summarizes the related work. Section III describes the development state of the minstrel robots we are working on. Next, Section IV explains how the posture classifier system has been developed, choosing among several supervised approaches and applying variable selection in order to learn a model that better fits our needs. Section V describes how the classifier has been integrated in the robot control architecture. Finally, section VI gives some conclusions and pinpoints future work.

## II. RELATED WORK

Human postures recognition is the closest example we have found related to the work presented here. A bunch of work can be found related to this problem, specially since the emergence of cheap 3D depth sensors like Microsoft's Kinect. On the one hand, there are depth images based approaches; Biswas and Basu [3] recognize gestures by extracting the variation of the body in depth images between each pair of consecutive frames and using a multiclass SVM to train the system. In [19] Shotton et al. propose a new method to predict the 3D positions of body joints using object recognition strategies and randomized decision forests. And Wang et al. [20] use features extracted from the ratio of the upper and lower human body and a LVQ neural network to recognize five human postures. On the other hand, among the skeleton information based approaches, Patsadu et al. [15] use a set of vectors of twenty body joint positions to recognize human gesture using various data mining classification methods, and then, they compare the performance of each method to find the optimal classifier. Reddy and Chattopadhyay [17] propose a method for human activity recognition based on skeleton joints and uses PCA and Statistical approaches for features extraction and SVM for activity classification. And Cicirelli et al. [6] use the quaternion features of the right shoulder and



elbow joints and uses different NNs to construct the models of 10 different gestures.

### III. PROBLEM CONTEXTUALIZATION

*Bertsolaritza* is a traditional Basque improvised singed poetry manifestation of the Basque Country that is still very much alive. From the point of view of social robotics, *bertsolaritza* offers a good scenario to develop new social behaviours. The *BertsoBot* project aims to develop troubadour robots. Beyond the automatic verse generation system, the goal is to join together the capabilities of autonomous robots to sense the environment and interact with it.

The *BertsoBot* system endows the robots with *bertsolari-s'* abilities, thus the system must follow the dynamic of a real *bertsolari-s* performance as troubadours do.

- 1) Await its turn sat on its chair.
- 2) Place itself in front of the microphone when required and listen to the exercise proposed by the emcee.
- 3) Compose and sing the verse to the public.
- 4) Observe and receive audience's feedback and react accordingly.
- 5) Go back to its resting location.

All these tasks are accomplished and managed by a ROS<sup>1</sup> based control architecture, composed by different behaviors or modules that makes the robot act in a consistent manner and resemble to a real *bertsolari*. Fig. 1 shows the global system architecture.

Currently, the "Robot State Controller" module –highlighted in the above figure with a red circle – works blindly, it assumes that the robot is in the correct state that enables it to fulfil the next possible action/movement/gesture set. The robot state is labelled according to the last action executed. That makes the system fragile in the sense that it is not able to recover from an unsuccessful movement or from an unknown initial state. Any self-aware robot, should be able to automatically recognize its body configuration.

The posture classifier system described in the next section aims to overcome this limitation taking into account that an important requirement is that the developed system should be able to give an accurate answer on the fly.

### IV. POSTURE CLASSIFIER SYSTEM

When the robot receives any order given by the emcee, it should be aware of its body posture. According to the flow of a performance two are the principal postures the robot can show: sit on a chair and stand up. But the robot often initializes in a typical comfort pose like crouched or sit on the floor. Thereby, four different body postures (shown in Figure 2) have been defined as states to be recognized.

We pretend the robot to be able to be self aware of its body position and thus, a skeleton based approach that uses proprioceptive sensors seemed more appropriate for the goal. We chose NAO's joint positions and rotation angles obtained from the robot's TF [8] (tree of coordinate frames),

<sup>1</sup>www.ros.org

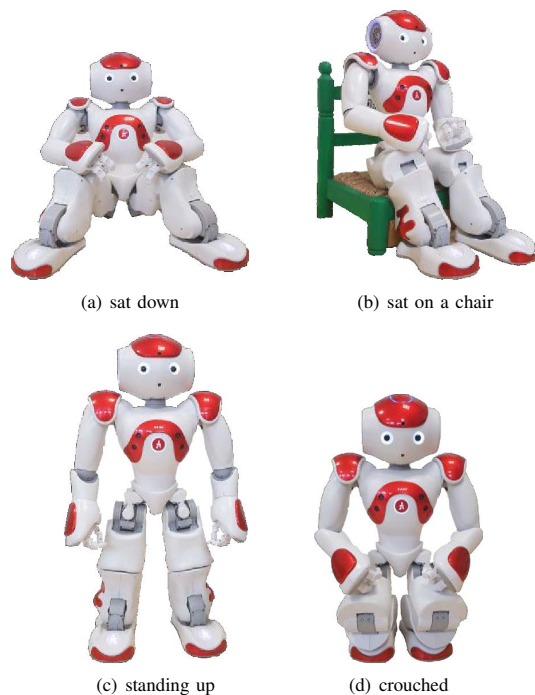


Fig. 2. Robot postures classes

which provides a standard way to keep track of coordinate frames and transforms data between different frames. Using this information we train our database and carried out the experimentation with different classifiers in order to get an appropriate model for the problem stated.

#### A. Posture Data set

NAO has 25 joints (see Figure 3), so there are 26 coordinate frames: one for each joint plus the root frame, named base link. As mentioned before, in our approach the features for the classification models are obtained from NAO's TF. In this transform tree the relationship between two coordinate frames is represented by 6 DOF composed by a point in the coordinate system,  $P(x, y, z)$ , and a quaternion,  $Q(\text{roll}, \text{pitch}, \text{yaw})$  for the orientation.

In order to collect the data, the robot has been manually set in each corresponding posture (standing up, sat on a chair, sat down and crouched) and afterwards softly moved with the stiffness off to get varying data. A feature vector of 150 ( $6 \times 25$ ) elements has been obtained from the transformation between the base link and each joint for each data entry. The data set contains 60 examples of each of four classes, hence 240 cases in total.

#### B. Classification Methods Used

This section gives a short introduction about the different classification methods considered in this work to build a model for posture recognition. Some of them are more efficient in the model generation process, while others are more effective in

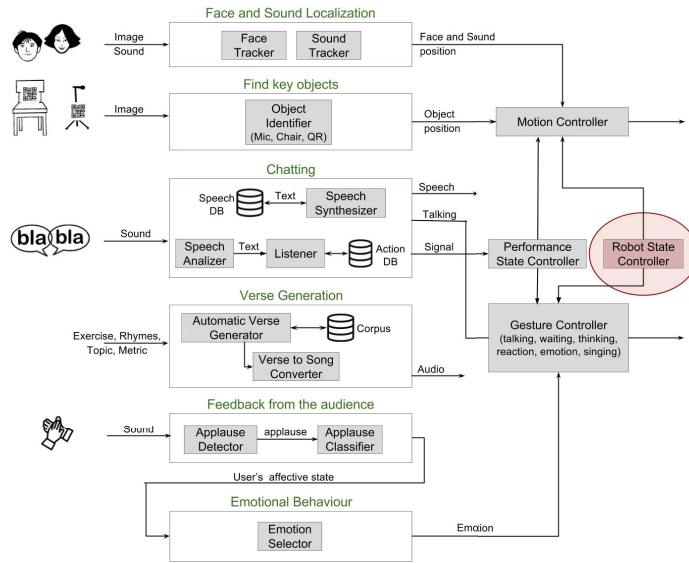


Fig. 1. BertsoBot global control architecture

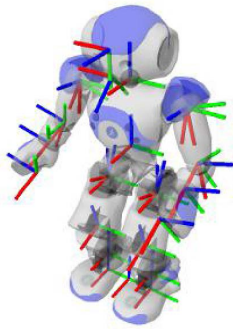


Fig. 3. NAO's tree of coordinate frames

the classification. In our case, the algorithm must be faster in the classification process, because the goal is to obtain the posture and detect if the robot is moving or not in “real-time”.

All the used classifiers have been trained and tested using WEKA [9], a well-known open source data mining tool.

*a) Decision trees:* A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. Several algorithms to generate such optimal trees have been devised, such as C4.5 [16] and CART (Classification and Regression Tree) [5].

J48 is the WEKA implementation of the wellknown C4.5

decision tree and the one used in the experiments.

*b) Naive Bayes algorithm:* A Naive Bayes (NB) classifier [21] is a simple probabilistic classifier based on applying Bayes’ theorem (from Bayesian statistics) with strong (naive) independence assumptions. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. An advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

WEKA’s NB version of the algorithm has been used with the default parameter values. It must be noted that NB’s underlying assumption of feature independence is not met in this case study.

*c) K-Nearest neighbours:* K-Nearest neighbours [1], is a distance based classification algorithm that stores all available cases and classifies new cases based on a similarity measure. This classifier is of lazy type, meaning that all the computational load resides in the classification of the instances, no training phase is applied. The classifier works directly on the data set, or training set. This implies that the entire data set has to be loaded into memory during each run of the classifier.

IBk is the implementation in WEKA of K-Nearest Neighbour algorithm, where K is the number of closest training examples.  $k = 1$  and  $k = 3$  have been tested during the experiments.

*d) Support Vector Machines:* A “Support Vector Machine” (SVM) [7] is a supervised machine learning algorithm which can be used for both classification or regression challenges. An SVM model is a representation of the examples as



points in space, mapped so that the examples of the separate categories are divided by a hyper plane that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. It supports a number of different kernels (hyper tangent, polynomial, and radial basis functions). The SVM learner supports multiple-class problems as well by computing the hyper plane between each class and the rest.

WEKA offers SMO, an implementation of SVM and it has been used with a polynomial kernel and default parameter set.

### C. Classification Results

Table I shows several performance measures for the classifiers previously described. Looking to the F-measure or F1 score, a measure of a test’s accuracy that consider the precision (P) and the recall (R) of the test to compute the score ( $2 * \frac{P * R}{P + R}$ ), J48 and IB3 classifiers give the best values. As mentioned before, a main requirement for the system being developed is the time needed to give a response. Taking into account the non-lazy nature of the classification tree it seems to better adjust to this requirement than the NN algorithm.

TABLE I  
EXPERIMENTS RESULTS WITH THE RAW DATA SET

	F-measure	Precision	Recall
J48	<b>0.971</b>	0.972	0.971
Naive Bayes	0.959	0.961	0.958
IB1	0.967	0.967	0.967
IB3	<b>0.971</b>	0.971	0.971
SMO	0.959	0.961	0.958

### D. Feature Subset Selection

The performance of the system in terms of accuracy largely depends on the set of features used. A high dimensionality can be a problem not only in terms of classification performance, it can also affect the time needed to give a response.

Feature Subset Selection (FSS) is the process of selecting the relevant subset of variables to construct the classifier. FSS algorithms can be broken up into Wrappers, Filters and Embedded. Filter methods, are based only on general features like the correlation with the variable to predict. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Embedded methods combine the advantages of both previous methods.

Wrapper methods have the ability to take into account feature dependencies. Thus we decided to try to reduce the feature vector dimensionality by applying a wrapper attribute selector to those outstanding classifiers identified in the previous step: J48 and IB3. We selected a greedy strategy as a search mechanism for feature subset selection and the assessment of each subset has been evaluated by 5-fold cross validation using the classification paradigm being evaluated in each case, i.e., J48 and IB3.

Looking at the results in table II IB3 slightly out stands J48’s performance with respect to the F-measure, but the decision

tree is a much faster estimator than the nearest neighbour algorithm.

TABLE II  
EXPERIMENTS RESULTS WITH WRAPPER

	F-Measure	Precision	Recall	Num. Attributes
J48	0.988	0.988	0.988	3
IB3	0.996	0.996	0.996	6

The J48 decision tree needs only three features out of the initial 150 to classify a pose: *RThigh\_yaw*, *RTibia\_roll* and *RAnklePitch\_x*). Moreover, it is step wise to export the model to NAO’s control architecture. Just a minimal set of if-then-else rules need to be implemented. Those and the non-lazyness nature of the decision tree are compelling reasons for opting for the J48 model as the final posture classifier for the *BertsoBot* system. Figure 4 shows the obtained tree.

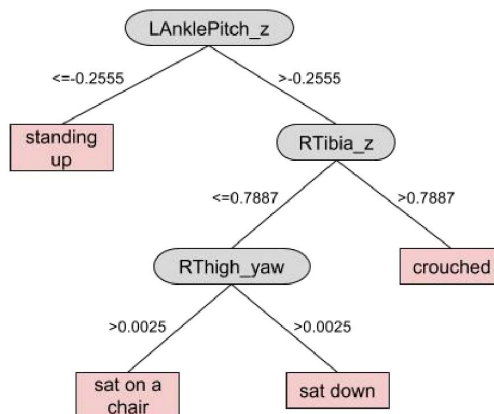


Fig. 4. Robot posture classification model defined as decision tree

## V. ROBOT SELF AWARENESS SYSTEM

Figure 5 shows the structure of the “Robot State Controller” after integrating the developed posture recognition system.

Notice that the posture classifier must be fed when the robot is stationary. To discriminate moving values, joint values at time  $t$  are subtracted from those at time  $t-1$  and any difference in any joint values implies that the robot is moving.

A ROS module has been implemented which endows the robot with self awareness capabilities that will help it to make a decision based on its body posture and state (moving or motionless). `nao_body_info` is the name of this package and it has two main purposes: predict the current posture of the robot and determine if it is moving or not.

`nao_body_info` package has been successfully integrated in the *BertsoBot* global system architecture. It outputs the current body information (posture and state) of the robot which serves as input of the gesture manager module. The current robot body information serves:

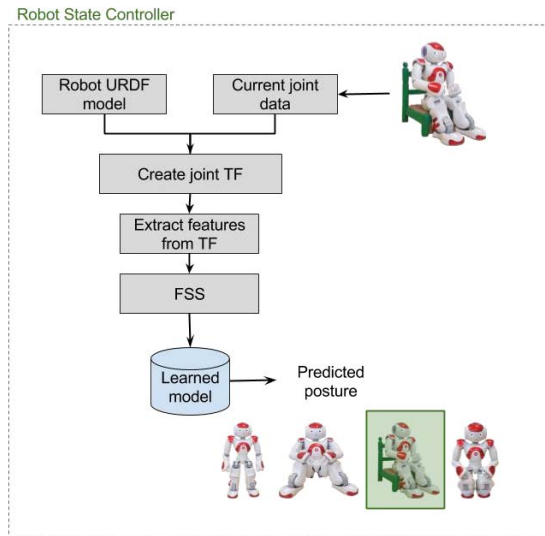


Fig. 5. Architecture of activity recognition system

- 1) as input to the gesture management module responsible of choosing the set of gestures the robot must show at each state (see [18] for more detailed information).
- 2) for ensuring that the robot is in the proper pose before performing an action such as reaching the micro or that the action has finished adequately – it is sat on the chair.

A short video<sup>2</sup> available at the RSAIT's youtube channel<sup>3</sup> demonstrates how the robot responds with a different action (movement) to the “stand up” order given verbally by the operator. NAO is able to recognise its initial pose – sat on a chair or sat on the floor, crouched or standing up– and decides the movements necessary to fulfil the goal accordingly. It nods to confirm it has understood the verbal command and tells the operator when it reaches the standing up position.

## VI. CONCLUSIONS AND FURTHER WORK

Throughout the paper we have described the process of developing a posture recognition system that endows the robot with a self awareness of its posture while acting as a minstrel robot. The system relies only on proprioceptive sensors (joint positions obtained from the servos). Several classifiers have been tested and the one that better fit the system requirements – the J48 decision tree– was selected and successfully integrated in the global robot control architecture.

Although the self awareness system is being used for entertainment purposes, it could also be helpful for any application which requires action planning that depends on the actor's body pose, for instance, to recover last body position when the

<sup>2</sup><https://www.youtube.com/watch?v=x88fK8IuYMc&t=4s>

<sup>3</sup><https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ>

robot falls, or to determine which was the last safe position before falling.

It remains as future work to extend the system to be able to recognize moving actions such as walking, sitting and so on. Increasing the self awareness would improve robot autonomy and thus, the robot behavior at the stage.

## ACKNOWLEDGMENT

This work has been partially supported by the Basque Government (IT900-16), the Spanish Ministry of Economy and Competitiveness MINECO/FEDER (TIN 2015-64395-R, MINECO/FEDER,EU). Author Rodriguez has received the UPV/EHU research Grant PIF13/104.

## REFERENCES

- [1] D.W. Aha, D. Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.
- [2] A. Astigarraga, E. Jauregi, E. Lazkano, and M. Agirrezabal. Textual coherence in a verse-maker robot. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*, pages 275–287. Springer, 2014.
- [3] K. K. Biswas and S. K. Basu. Gesture recognition using microsoft kinect®. In *Automation, Robotics and Applications (ICARA), 2011 5th International Conference on*, pages 100–103. IEEE, 2011.
- [4] J. Bongard, V. Zykov, and H. Lipson. Resilient machines through continuous self-modeling. *Science*, 314(5802):1118–1121, 2006.
- [5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.
- [6] G. Cicirelli, C. Attolico, C. Guaragnella, and T. D’Orazio. A kinect-based gesture recognition approach for a natural human robot interface. *International Journal of Advanced Robotic Systems*, 12(3):22, 2015.
- [7] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [8] T. Foote. TF: The transform library. In *Technologies for Practical Robot Applications (TePRA), 2013 IEEE International Conference on, Open-Source Software workshop*, pages 1–6, April 2013.
- [9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, R. Reutemann, and I. H. Witten. The WEKA data mining software: An update.
- [10] J. W. Hart and B. Scassellati. Mirror perspective-taking with a humanoid robot. In *AAAI*, 2012.
- [11] C. Von Hofsten. An action perspective on motor development. *Trends in cognitive sciences*, 8(6):266–272, 2004.
- [12] E. O. Johnson and P. N. Soucacos. Proprioception. *International Encyclopedia of Rehabilitation*, 2010.
- [13] H. Lagercrantz and J.P. Changeux. The emergence of human consciousness: from fetal to neonatal life. *Pediatric Research*, 65(3):255–260, 2009.
- [14] P. Michel, K. Gold, and B. Scassellati. Motion-based robotic self-recognition. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 3, pages 2763–2768. IEEE, 2004.
- [15] O. Patsadu, C. Nukoolkit, and B. Watanapa. Human gesture recognition using kinect camera. In *Computer Science and Software Engineering (JCSE), 2012 International Joint Conference on*, pages 28–32. IEEE, 2012.
- [16] J. Ross Quinlan. C4. 5: Programming for machine learning. *Morgan Kaufmann*, 38, 1993.
- [17] V. R. Reddy and T. Chattopadhyay. Human activity recognition from kinect captured data using stick model. In *International Conference on Human-Computer Interaction*, pages 305–315. Springer, 2014.
- [18] I. Rodriguez, A. Astigarraga, T. Ruiz, and E. Lazkano. Singing minstrel robots, a means for improving social behaviors. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 2902–2907. IEEE, 2016.
- [19] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.
- [20] Wen-June Wang, Jun-Wei Chang, Shih-Fu Huang, and Rong-Jyue Wang. Human posture recognition based on images captured by the kinect sensor. *International Journal of Advanced Robotic Systems*, 13(2):54, 2016.
- [21] H. Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.

## 12.5 On how self-body awareness improves autonomy in social robots

**Title:** On how self-body awareness improves autonomy in social robots

**Authors:** I. Rodriguez, J. M. Martínez-Otzeta, E. Lazkano, T. Ruiz, B. Sierra

**Conference:** International Conference on Robotics and Biomimetics (ROBIO)

**Publisher:** IEEE

**DOI:** 10.1109/ROBIO.2017.8324661

**Year:** 2017

# On how self-body awareness improves autonomy in social robots

I. Rodriguez<sup>1</sup>, J. M. Martínez-Otzeta<sup>1</sup>, E. Lazkano<sup>1</sup>, T. Ruiz<sup>2</sup> and B. Sierra<sup>1</sup>

**Abstract**—Just as humans show consciousness of their body, social robots, in the way to be truly autonomous need to be aware of their body posture. Feasible gestures, moves and actions depend on the current body posture. The work developed in this paper aims to empirically show how self configuration recognition augments the degree of autonomy of a robot in the context of entertainment robotics. The integration of a classification tree for body posture identification based on data acquired from proprioceptive sensors of a NAO robot allows to interact with the robot in a more flexible and persistent manner. As a result, the robot shows a more sound behavior and greater degree of autonomy. Moreover, even if the body-awareness has been developed for minstrel robots, its application can be generalized to other contexts.

## I. INTRODUCTION

According to the Merriam-Webster dictionary, autonomy is the quality or state of being self-governing. However, the concept of autonomy in robots goes further and comprises many qualities such as long term functioning, adaptability, learning capabilities, operation with little human intervention, self detection of errors, etc.

There is agreement in the robotics community that autonomy is not a yes or not property; the degree of autonomy of a robot is a characteristic that ranges from no autonomy to high autonomy. In that vein, Patrick Rau et al. [18] define autonomy as the degree to what a robot can act on its own accord. Moreover, the definition of a robot's autonomy level depends on the task and environment of the robot. The autonomy of robots that operate in isolation or that of rescue robots for instance differs from that of social robots.

We found few attempts to define a taxonomy for the autonomy degree in robots. Most noteworthy, the guideline for categorizing the autonomy level of a robot (LORA) proposed in [2] in the context of HRI and focusing on service robots. The autonomy level is usually associated with the need for human intervention and thus, robots that need to be operated by humans to perform well have low autonomy and robots able to sense-plan-act with minimal human input are categorized as highly autonomous. As Patrick Rau et al. underline in [18], this autonomy level classification does not suit social robots because the robot performance is directly linked to the adequate interaction with humans. Most socially interactive robots do not have yet the ability to work unattended for extended periods of time. In fact, most of

them are either remotely operated or follow a very specific set of rules [7].

This paper contributes to show that body self awareness improves autonomy in social robots. As mentioned before, the autonomy level is directly linked to the robot/environment system. We focus on entertainment robots that must act as minstrels. Experimental results show that the interaction with the emcee is highly enriched if robots are aware of their body postures.

The rest of the paper is organized as follows: Section II reviews robot self awareness. Next, Section III describes the context in which the robot should actuate because autonomy is directly linked to the robot and the task. Section IV explains how the body posture awareness system has been implemented and the changes introduced in the robot control architecture are detailed in Section V. The overall system is evaluated in Section VI and the last Section gives conclusions and points to further work.

## II. SELF AWARENESS

According to Lagercrantz et al. [13] a simple definition of consciousness is the sensory awareness of the body, the self, and the world. The first thing children perceive is their own body, which serves as a means of interaction with others and the environment. Thanks to her/his body, the child experiences different sensations, mobilizes and learns [10]. Recognizing oneself, although it seems easy, at least requires to be conscious of one's body and one's actions. The five senses (sight, hearing, smell, touch and taste) are the traditionally recognized methods of perception responsible for our interaction with the external world. But additionally to the exteroceptive sensors, we have senses that are responsible for our internal functioning. That fundamental internal sensory system, called proprioception, provides feedback about the status of the body internally [11]. Thanks to muscle spindles, which detect changes in the muscle and signal the angle of related joints, we get information about limb positions, and we realize our body's position.

Just as humans show consciousness of their body, social robots, in the way to be truly autonomous, should also be able to recognize their own configuration. Various social robotic projects and researchers address the problem of user awareness. For instance, Anki's Cozmo<sup>1</sup> is a tiny robot with impressive body expression that incorporates face recognition capabilities for owner identification purposes. Bergner et al. [3] are working in an artificial skin with multiple sensors to cover the whole body of a robot with the aim of acquiring

<sup>1</sup>I. Rodriguez, J. M. Martínez-Otzeta, E. Lazkano and B. Sierra are with Faculty of Informatics, Computer Sciences and Artificial Intelligence, University of the Basque Country (UPV/EHU), 20018 Donostia [igor.rodriiguez@ehu.eus](mailto:igor.rodriiguez@ehu.eus)

<sup>2</sup>T. Ruiz is with Faculty of Informatics, Computer Architecture and Technology, University of the Basque Country (UPV/EHU), 20018 Donostia

<sup>1</sup><http://www.anki.com/en-us/cozmo/cozmo-tech>

human awareness and enhancing HRI. Besides, Lanillos et al. [14] address the problem of self-perception using a hierarchical Bayesian computational model that integrates proprioceptive and tactile cues with visual cues to allow the robot distinguish between in-body and out-body elements in the scene.

Several robotic systems can be considered as self-aware systems to some degree, being able to recognize themselves in the mirror [9], or being aware of their motion [16]. In [15] H. Lipson shows how a walking robot can learn the walking patterns of the legs by interacting with the environment. Finally, Bongard et al. [5] propose a system able to change their own models of their physical embodiment.

Social robots with humanoid bodies need to be aware of their body posture. Feasible gestures, moves and actions depend on the current body posture. Multiple works can be found related to human posture recognition: approaches based on depth images [4][22][23] or approaches that rely on skeleton information [17] [19] [6], to mention some. We tackled the problem in a rather different way. Taking into account only internal sensory receptors, the posture recognition system developed endows the robot NAO with the ability to know its body posture without the aid of visual information. The posture recognizer relies only in proprioceptive information and a supervised classification approach is used to acquire the model. It is a more engineering approach to develop body awareness, but it is indeed a way to increase the self awareness of the robot. According to the posture detected, the robot knows what type of movements or actions can perform and thus, the interaction capability is enriched due to an increase of the autonomy level.

### III. BERTSOBOT

As mentioned before, we focus on minstrel robots that sing improvised poetry in Basque, *Euskara*, the language of the inhabitants of the Basque Country. *Bertsolaritza*, Basque improvised contest poetry, is one of the manifestations of traditional Basque culture that is still very much alive.

Events and competitions are very common in the Basque Country, which usually consist of a rather formal flow of poetry recitations, *bertso-s*. In such performances, several verse-makers, named *bertsolari-s*, compete with each other singing improvised verses about topics or prompts which are given to them by an emcee (theme-prompter). They compose the verses on the fly, normally in less than one minute, and sing a poem along the pattern of a prescribed verse-form that also involves a rhyme scheme. Melodies are chosen from among hundreds of tunes.

The aim of any *bertsolari* when she/he sings an improvised verse is to convey a message, according to the requirements imposed by the emcee, while entertaining the public. But the interaction with the environment – stand in front of the microphone, obtain rhymes given by the emcee, perceive audience’s reaction – is as important as the composition of the verse.

We are working on a project named *BertsoBot* which aims to develop social minstrel robots for entertainment [1] [20].

The *BertsoBot* system endows the robots with some of the *bertsolari-s*’ capabilities that allow these social robots to take part in a *bertsolari-s* performance.

For that purpose, the system must follow the dynamic of real events, *bertso-saio-s*, as troubadours do:

- 1) Wait for its turn, sit on a chair or stood up beside other troubadours.
- 2) When its turn arrives, place itself in front of the microphone and listen to the exercise proposed by the emcee.
- 3) Compose and sing the verse to the public.
- 4) Observe and receive audience’s feedback and react accordingly.
- 5) When finished, go back to its sitting place.

These five steps summarize the basic flow of an event. Each *bertsolari* is called several times by the emcee, either alone or together with one or more fellows, and each time they can be mandated to sing several verses, i.e. steps 2 to 4 are looped.

All these tasks are accomplished and managed by a ROS<sup>2</sup> based control architecture, composed by different behaviors or modules, that makes the robot act in a consistent manner and resemble to a real *bertsolari*. Fig. 1 shows the initial control architecture developed and used in the public performances made up to September 2016. See some of these performances videos at RSAIT’s youtube channel<sup>3</sup>.

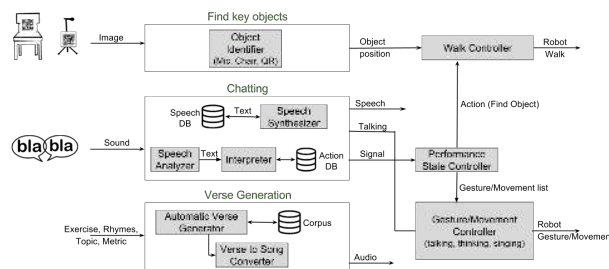


Fig. 1. Original system architecture

The verse is composed and sung by the “Verse Generation” process, the “Chatting” behaviors allow the interaction with the emcee, while “Find key objects” provides the robot with necessary skills to interact with environmental key objects, such as the microphone and its resting place. These interactions, usually executed as walking actions, are managed by the “Walk Controller”. The robot body expression is managed by the “Gesture controller” which decides the gestures to be applied from the appropriate gesture set at each state according to the “Performance State Controller”. The gesture sequence is randomly selected at each time so that the robot doesn’t show the same body behavior one and again (see [20] more details).

<sup>2</sup>www.ros.org

<sup>3</sup>https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ

### A. System Limitations

The performance was totally driven by the verbal commands given by the emcee that acted as a sequencer. These commands were translated into signals by the “Chatting” module and given as input to the “Performance State Controller”. These signals are associated to different orders, such as wake up, rest, sit down on chair, sit down on the floor, find micro, get exercise, sing a verse, find chair and so on.

Thus the flow of an event completely relied on the “Performance State Controller”, a finite state automaton that operated in an open-loop. The initial setup of the system should always be the same (the robot started sat on the chair). Any discordance between the real state and the sequence of actions to be executed produced an undesirable global behavior and required the intervention of the operator and the interruption of the performance.

## IV. BODY POSTURE AWARENESS

According to the flow of a performance two are the principal postures the robot can show: sat on a chair and stood up (see Section III). But the robot often initializes in a typical comfort pose like crouched or sat on the floor. Thereby, four different body postures (sat down on the floor, sat on a chair, sat on a chair, stood up and crouched) have been defined as body states to be recognized (see Fig. 2).

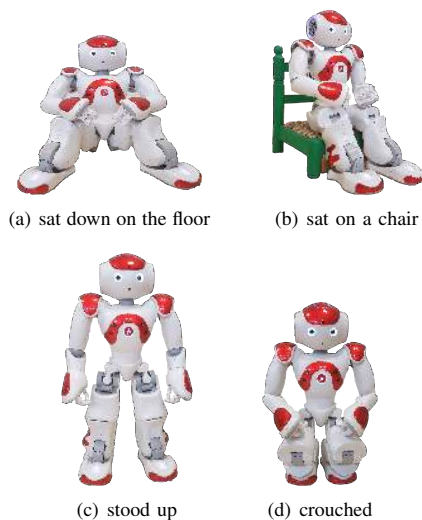


Fig. 2. Posture classes

The taken approach relies on a skeleton based approach that uses proprioceptive sensors. Given that NAO has 25 joints and each joint is characterized by six values, 150 values suffice to define the posture of the robot.  $P(x, y, z)$  positions and  $Q(roll, pitch, yaw)$  Euler angles are obtained and transformed to the robot’s *base.link* frame using TF library [8], which provides a standard way to keep track of coordinate frames and transforms data between different frames.

We can think of every posture of the robot as associated to a vector of those 150 values, and the problem of recognizing

the posture is one of classification: different postures are associated to different vectors and given a vector, the robot position has to be inferred.

This problem lies in the field of supervised learning, where from a training set of known instances a model is built to predict the correct class of new presented instances. The first step, then, is to get some known instances to which apply machine learning methods. This has been achieved recording the joint positions and Euler angles of the NAO robot while in several different variants of the robot postures. This gave us a total of 3332 training instances distributed as shown in Fig. 3. Currently, we are using a J48 decision tree machine learning paradigm because we are interested in the easy implementability and explicability of the generated model though other approaches have been tested with no better results [21]. 10-fold crossvalidation results are shown in Table I. The obtained accuracy is close to 100%.

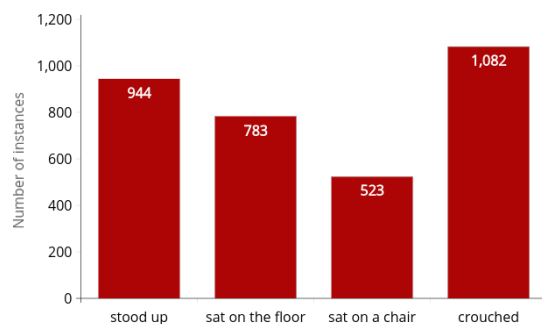


Fig. 3. Data distribution for training

TABLE I  
RESULTS OF 10-FOLD CV FOR THE J48 DECISION TREE

	mean	std
accuracy	0.9997	0.0009
jaccard similarity	0.9997	0.0009
zero one loss	0.0003	0.0009

The learned tree structure can be seen in Fig. 4. Only 5 from 150 features turned out to be relevant for the classification process: right tibia pitch and yaw, right thigh yaw, left thigh pitch and left elbow yaw.

It is also not difficult to interpret the generated tree. In a first look it is seen that the values associated to the joints at the tibia, thigh and elbow are used to discriminate between postures. This is not contradictory with which a person could think of.

Once the model is built, the posture recognition process is quite simple. It can be summarized into two main steps as shown in Fig. 5. In the initial step, required features are obtained from the robot’s TF. And then, the decision tree learned before is applied to the extracted data, which returns as output the predicted posture of the robot.



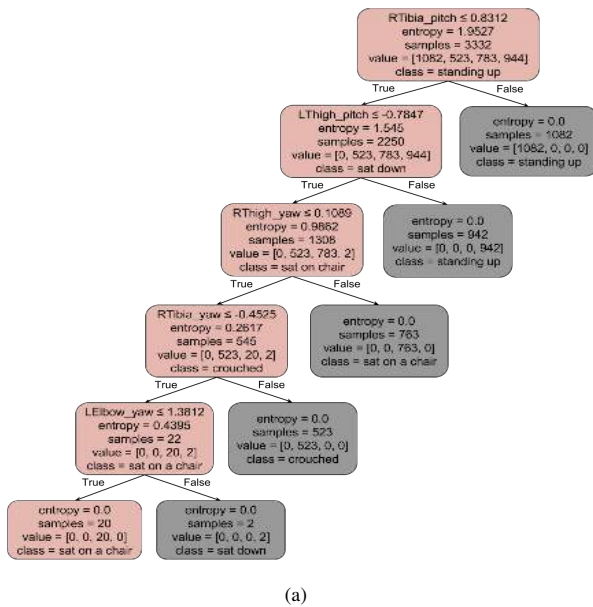


Fig. 4. Entropy decision tree (Given for repeatability purposes)

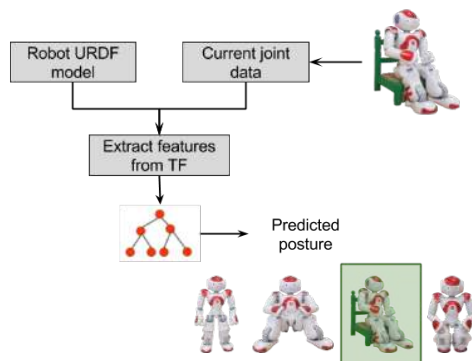


Fig. 5. Posture selection mechanism

## V. MODIFICATIONS IN THE CONTROL ARCHITECTURE

Fig. 6 shows the schema of the new system architecture. Several new modules have been added to enrich the sociability of the robot. On the one side, “Face and Sound Localization” allows the robot to track the emcee’s face and redirect its body towards the sound source to give the audience the illusion that the robot pays attention to what the emcee is saying. On the other side, the audience applause, captured and classified by the “Feedback from the audience” behavior, affect the robot’s emotional state. Besides, the “Emotional Behavior” module pretends to adapt the body language expression to the response received from the audience (see [12]).

The signals come from the “Chatting” module and feed the goal of the “Performance State Controller” as before. But the main concern of the work described here resides on the “Body awareness” module (red colored rectangle in Fig. 6) and its coordination with the existing behaviors. This

“Body Awareness” module decomposes the given order into an action or sequence of actions selected according to the current posture of the robot and verifies whether the order has been successfully completed or not. The module is composed by two new nodes:

- The “Body Posture Recognizer”: implements the decision tree previously shown in Fig.4. It outputs the current body posture taking the robot joint information as input.
- The “Action Decomposer”: receives the order and translates it into an action or sequence of actions to perform according to the current posture of the robot. Each action emerges in a gesture or a movement, and must end in a specific posture. After the execution of the action, it verifies that the gesture/movement ends successfully and it checks whether the actual posture of the robot corresponds to the expected one. It also controls the postures in which walking is possible. For instance, if the robot needs to reach the microphone, it needs to stand up before it starts to walk. This behavior verifies this condition and commands the proper movements (depending upon the current posture) to the “Gesture/Movement Controller” to achieve the desired posture.

Again, the robot body expression is managed by the “Gesture/Movement Controller”, which decides the gestures to be applied from the appropriate posture/gesture set at each state. But now, the correct gesture set is chosen depending on the information provided by the “Body Awareness” and thus, only gestures feasible at the real body posture are executed.

## VI. SYSTEM EVALUATION

We intend to show that robot posture self awareness helps to increase the autonomy level and at the same time, facilitates the interaction with the robot, showing a better behavior performance in the context of singing minstrel robots, i.e. *bertsolari* robots. As mentioned in the introduction, autonomy is a gradual property not easy to be measured neither quantitatively nor qualitatively. Thus, we decided to empirically show the improvements introduced in the system by incorporating the “Body Awareness” module. Two experiments have been performed and recorded:

- 1) Experiment 1: the robot receives a “stand up” order from every different pose. The video<sup>4</sup> shows how the robot adjusts the movement to be executed depending on its current body posture to obey the order given by the human. NAO nods to confirm it has understood the command and when it reaches the desirable final posture it says so.
- 2) Experiment 2: the *bertsolari*-s are not always called to approach the microphone directly while sitting on the chair, they can be required to first stand up and listen to the exercise, before they approach the microphone to sing. In this second video<sup>5</sup> the robot is told to reach

<sup>4</sup><https://www.youtube.com/watch?v=x88fK8IuYMc&t=25s>

<sup>5</sup><https://www.youtube.com/watch?v=0uax6qilK30>

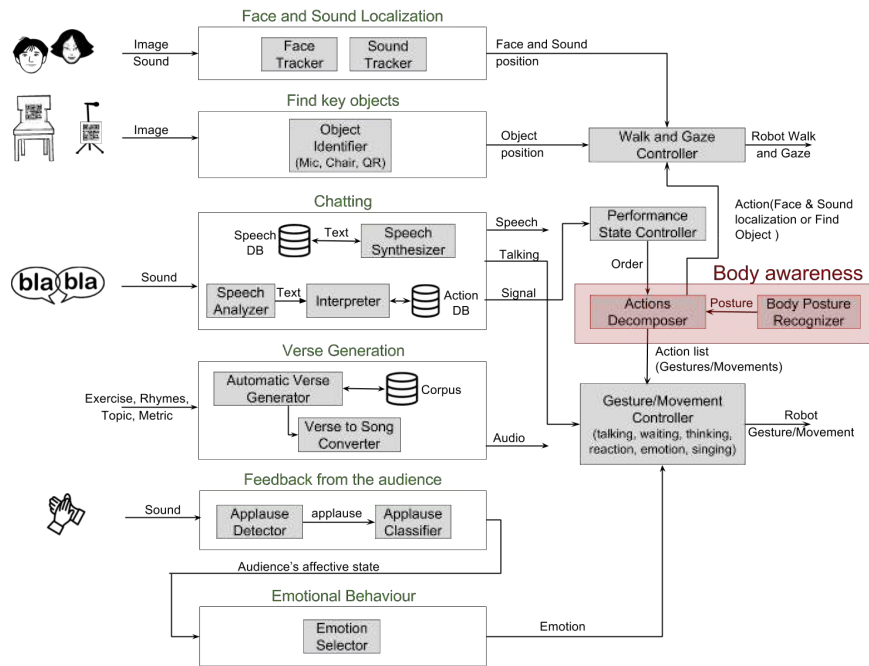


Fig. 6. Overall architecture

the microphone, again from different initial postures. The complexity of this order (as that of sitting on the chair) is higher because the reference element does not need to be in the robot's current field of view.

## VII. CONCLUSIONS AND FURTHER WORK

This work describes how robot body posture self awareness improves the autonomy of the system in the context of entertainment robots. The obtained posture classification system is provided and thus, the module can be replicated in different NAO robots. Moreover, even if the body-awareness has been developed for minstrel robots, its application can be generalized to other contexts.

The improvements introduced in the overall system are three-fold:

- 1) The robot gestures always match the gestures allowed by its current body position and thus, no weird behavior occurs. The differences between the two versions of the architectures can be appreciated in this short video<sup>6</sup>.
- 2) There is no need to execute the same sequence of actions during the performance. The emcee is free to change the flow in real time and the initial setup is unnecessary. The robot adapts the actions to perform according to its current state independently of the initial posture.
- 3) More complex actions can be defined that comprise several subgoals. The emcee does not need to worry

about each next step anymore.

Despite the difficulty of measuring the autonomy level of a social robot, we believe that these improvements increase the degree of autonomy of the system. The "Body Awareness" module permits to derive robot's action algorithmically instead of being prescribed by a human, adapting those actions/gestures to whatever its body posture is. As a consequence, it allows for a more flexible and persistent interaction and improves the overall behavior of the robot. In the context we are working on, these two features, flexibility to modify the flow of the performance in real time and persistent and robust interaction contribute to autonomy.

In the particular setting presented the "Body Posture Recognizer" has to distinguish only four postures and thus, it could be argued that some of them have very different knee joint positions and that they could be classified in an ad-hoc manner. However, the aim is to develop a method that will serve us for other postures as well, i.e., to enrich the posture recognizer with more poses. It remains as immediate further work to do the integration of fall detection and fall recovery. Adding this new pose to the set of postures will allow to recover from badly performed moving actions and increase one step further the global autonomy.

## ACKNOWLEDGMENT

This work has been partially supported by the Basque Government (IT900-16), the Spanish Ministry of Economy and Competitiveness MINECO/FEDER (TIN 2015-64395-R, MINECO/FEDER,EU). Author Rodriguez has received the UPV/EHU research Grant PIF13/104.

<sup>6</sup><https://www.youtube.com/watch?v=ZiDaSRloMWg>



## REFERENCES

- [1] A. Astigarraga, E. Jauregi, E. Lazkano, and M. Agirrezabal. Textual coherence in a verse-maker robot. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*, pages 275–287. 2014.
- [2] J. Beer, A.D. Fisk, and W. A. Rogers. Toward a framework for levels of robot autonomy in human-robot interaction. *Human-Robot Interaction*, 3(2):74–99, 2014.
- [3] F. Bergner, E. Dean-Leon, and G. Cheng. Event-based signaling for large-scale artificial robotic skin - realization and performance evaluation. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 4918 – 4924. IEEE/RSJ, 2016.
- [4] K.K. Biswas and S.K. Basu. Gesture recognition using Microsoft kinect®. In *Automation, Robotics and Applications (ICARA), 2011 5th International Conference on*, pages 100–103. IEEE, 2011.
- [5] J. Bongard, V. Zykov, and H. Lipson. Resilient machines through continuous self-modeling. *Science*, 314(5802):1118–1121, 2006.
- [6] G. Cicirelli, C. Attolico, C. Guaragnella, and T. D’Orazio. A kinect-based gesture recognition approach for a natural human robot interface. *International Journal of Advanced Robotic Systems*, 12(3):22, 2015.
- [7] B.J. Dunstan, D. Silvera-Tawil, J. T.K.V. Koh, and M. Velonaki. *Cultural Robotics*, chapter Cultural Robotics: robots as participants and creators of culture, pages 3–13. Springer International Publishing, 2016.
- [8] T. Foote. TF: The transform library. In *Technologies for Practical Robot Applications (TePRA), International Conference on, Open-Source Software workshop*, pages 1–6. IEEE, April 2013.
- [9] J. W. Hart and B. Scassellati. Mirror perspective-taking with a humanoid robot. In *26th AAAI Conference on Artificial Intelligence*, 2012.
- [10] C. Von Hofsten. An action perspective on motor development. *Trends in cognitive sciences*, 8(6):266–272, 2004.
- [11] E. O. Johnson and P. N. Soucasos. Proprioception. *International Encyclopedia of Rehabilitation*, 2010.
- [12] F. Kraemer, I. Rodriguez, O. Parra, T. Ruiz, and E. Lazkano. Minstrel robots: body language expression through applause evaluation. In *International Conference on Humanoid Robots (Humanoids)*, pages 332–337, Cancun (Mexico), November 2016. IEEE-RAS.
- [13] H. Lagercrantz and J.P. Changeux. The emergence of human consciousness: from fetal to neonatal life. *Pediatric Research*, 65(3):255–260, 2009.
- [14] P. Lanillos, E. Dean-Leon, and G. Cheng. Yielding self-perception in robots through sensorimotor contingencies. *IEEE transactions on cognitive and developmental systems*, 9(2):100–112, 2017.
- [15] H. Lipson. Building self awareness robots. [https://www.ted.com/talks/hod\\_lipson\\_builds\\_self\\_aware\\_robots](https://www.ted.com/talks/hod_lipson_builds_self_aware_robots), Accessed 2017-06-22.
- [16] P. Michel, K. Gold, and B. Scassellati. Motion-based robotic self-recognition. In *International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2763–2768. IEEE/RSJ, 2004.
- [17] O. Patsadu, C. Nukoolkit, and B. Watanapa. Human gesture recognition using kinect camera. In *International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 28–32. IEEE, 2012.
- [18] P.L. Patrick Rau, Y. Li, and J. Liu. Effects of a social robot’s autonomy and group orientation on human decision-making. *Advances in Human-Computer Interaction*, 2013.
- [19] V. R. Reddy and T. Chattopadhyay. Human activity recognition from kinect captured data using stick model. In *International Conference on Human-Computer Interaction*, pages 305–315. Springer, 2014.
- [20] I. Rodriguez, A. Astigarraga, T. Ruiz, and E. Lazkano. Singing minstrel robots, a means for improving social behaviors. In *International Conference on Robotics and Automation (ICRA)*, pages 2902–2907. IEEE, 2016.
- [21] I. Rodriguez, A. Astigarraga, T. Ruiz, and E. Lazkano. Body self-awareness for social robots. In *International Conference on Artificial Intelligence, Robotics and Optimization (ICCAIRO)*. CPS, 2017.
- [22] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.
- [23] Wen-June Wang, Jun-Wei Chang, Shih-Fu Haung, and Rong-Jyue Wang. Human posture recognition based on images captured by the kinect sensor. *International Journal of Advanced Robotic Systems*, 13(2):54, 2016.

## 12.6 Adaptive Emotional Chatting Behavior to Increase the Sociability of Robots

**Title:** Adaptive Emotional Chatting Behavior to Increase the Sociability of Robots

**Authors:** I. Rodriguez, J. M. Martínez-Otzeta, E. Lazkano, T. Ruiz

**Conference:** International Conference on Social Robotics (ICSR)

**Publisher:** Springer

**DOI:** 10.1007/978-3-319-70022-9\_66

**Year:** 2017

# Adaptive Emotional Chatting Behavior to Increase the Sociability of Robots

Igor Rodriguez<sup>1</sup>( ), José María Martínez-Otzeta<sup>1</sup>, Elena Lazkano<sup>1</sup>,  
and Txelo Ruiz<sup>2</sup>

<sup>1</sup> Computer Sciences and Artificial Intelligence, University of Basque Country  
(UPV/EHU), Manuel Lardizabal 1, 20018 Donostia, Spain  
`igor.rodriguez@ehu.eus`

<sup>2</sup> Computer Architecture and Technology, University of Basque Country  
(UPV/EHU), Manuel Lardizabal 1, 20018 Donostia, Spain

**Abstract.** Emotion expression is one of the characteristics that make us social beings. It is one of the main forms, along with oral and written language, that gives us a glimpse into the inner mental state of another individual. One of the aims of social robotics is the effortless communication between humans and robots. To achieve this goal, robotic emotional expression is a key ability, as it offers a more natural way to interact in a human-robot environment. In this paper a system to express the emotional content of a spoken text is presented. Head and arms movements, along with eye LED lighting and voice intonation are combined to make a humanoid robot express the sadness-happiness emotion continuum. The robot is able to express the emotional meaning of texts in English, Spanish and Basque languages.

**Keywords:** Emotion expression · Humanoid robot · Sentiment analysis · Body language

## 1 Introduction

Until recently it was unthinkable that someday humans and robots would be able to live together sharing the same space. But nowadays, robots start to become companions or co-workers of humans, opening an important research domain to build social robots that are able to naturally interact with us. Verbal communication is the most natural communication way that humans use for social interaction, but it is non-verbal communication what really helps us to understand sociability [1]. Body language is an important mean of communication; the gestures, postures and movements of the body and face are used to convey information about the emotions and thoughts of the sender while supporting verbal communication. In other words, body language is the key to express emotions.

Social robots must be expressive in a human-like way in order to be socially accepted. Robots that show high facial and body expressiveness have already been successfully developed. For instance, Nexi [2], developed by MIT Media

Lab, shows a wide assortment of facial expressions to communicate with people in human-centric terms. Kismet [3] is another robotic head that represents itself a milestone as how the human voice and facial features affect expressiveness. However, the advent of humanoid robots has prompted researchers to investigate and develop body language expression in robots. Softbank's NAO and Pepper<sup>1</sup> are two platforms that fit for that purpose thanks to their human-like shape and high expression capabilities.

The goal of this work is to endow robots with the ability to adapt their way to express different emotions according to the sentiment of the speech. The adaptive emotional system that we propose combines head position, arms gesticulation, eye LED lighting and voice intonation, making a humanoid robot able to express an emotion in the sadness-happiness continuum. An experiment has been designed where people is faced to a NAO robot with different body language, facial expressions and voice intonations according to a predefined inner emotional state. They have been asked about their opinion relating to the different options displayed for each type of communication way.

## 2 Related Work

In the Virtual Agent community a lot of research on emotional behavior generation has been conducted over the last 20 years. Although this is not directly transferable to robots with their reduced expressive means, there are great similarities in the overall architectures [4–6].

According to Breazeal [7], sociable robots are socially intelligent robots in a human-like way, and interaction with them is like interacting with people. As social and companion robots come to the market, the need to develop robotic systems with more complex behaviors is increasing. In recent years a lot of effort has been put in trying to make those behaviors convey sentiment. Several work can be found related to robots showing emotions through facial expression. Johnson et al. [8] investigate how LED patterns around the eyes of Softbank's NAO robot can be used to imitate human emotions. Some more recent work like Paradedá's [9] shows that the level of trust that a human being displays during an interaction with a robot is highest when the robot starts with small talk and expresses facial expression in the same direction of the storytelling expected emotion. In [10] a research is shown about the capabilities of a low-resolution RGB-LED display in the context of artificial emotions. They focus on four human emotions: happiness, anger, sadness and fear, and work with colors and dynamic light patterns which are supposed to evoke various associations.

Relevant work can also be found related to the body expression. Beck et al. [11] investigate the creation of an Affect Space (*Valence*, *Arousal* and *Dominance* (VAD)) for the generation of emotional body language to be displayed by robots; they assessed the effect of varying a robot's head position on the interpretation of predefined emotional key poses. Later, these authors

---

<sup>1</sup> <https://www.ald.softbankrobotics.com/en/robots>.

made a similar study [12] in which they tested children's ability to recognize the emotional body language displayed by a humanoid robot. Tielman et al. [13] define a model for a expressive behavior of the NAO robot in which *Valence* and *Arousal* values are influenced by the emotional state of its interaction partner and emotional occurrences while interacting with its environment. NAO expresses these emotions through its voice, eye color, and predefined posture and gestures. McColl et al. [14] have developed emotional body language for Brian 2.0 robot using a variety of body postures and movements identified in human emotion research. In another approach, Bretan et al. [15] explore the affective expression capabilities of Shimi robot, which has a small number of degrees of freedom, non-humanoid design and no face. Through several experiments they show that this kind of robots can also be emotionally expressive.

### 3 Emotion Expression Behavior

In order to make the robots able to express the emotional content of a spoken text, two main tasks are required: extract the emotion from the text and translate it to a robot expression. This process can be summarized into three main steps:

1. Text sentiment analyzer assesses the sentiment of the text and outputs a descriptor with information about the polarity of the sentiment (negative/neutral/positive) and a numerical value of the emotion in *Valence-Arousal-Dominance* space.
2. Emotion selector analyses the incoming descriptor and according to the polarity and VAD values it outputs the assessed emotion of the speech. In our approach only happiness and sadness emotions have been considered.
3. Emotion translator displays the emotion selected in the previous step by means of body expression, facial expression, and voice expression.

#### 3.1 Text Sentiment Analyzer

Sentiment analysis is the research field related to the analysis of people's opinions, sentiments, evaluations, attitudes, and emotions from written language. The theory of basic emotions states that emotions can be divided into discrete and independent categories [16]. On the other side, dimensional affective models regard affective experiences as a continuum of highly interrelated and ambiguous states. Emotions are described as linear combinations of *Valence-Arousal-Dominance* (VAD). *Valence* defines how positive or negative the stimulus is, *Arousal* specifies the level of energy and *Dominance* (also referred to as *Stance*) defines how approachable the stimulus is. The basic purpose of sentiment analysis is to extract the polarity (negative/neutral/positive) of a given text, but more advanced sentiment analyzers appraise the emotional content from the text as emotional state (happiness, anger, sadness, etc.).

The main task of the sentiment analyzer module is to carry out the emotion extraction process. It must provide the polarity of a text according to the

negative-neutral-positive category system, and also a numerical value of the emotion in the *Valence-Arousal-Dominance* space. Although our primary interest is a working tool in Basque language, we also want the system work with texts in Spanish and English.

The developed module combines two open-source tools in order to obtain the polarity and the emotion from the text: EliXa [17] and MixedEmotions toolbox<sup>2</sup>. EliXa is an aspect based sentiment analysis platform for text in English, Spanish and Basque. It assesses the polarity of a text according to a three category classification model: negative, neutral and positive. It returns one of these classes but without any confidence level. On the other hand, MixedEmotions platform is a Big Data Toolbox for multilingual and multimodal emotion extraction and analysis. It follows the *Valence-Arousal-Dominance* model and it is available in English and Spanish. A confidence level for each of six basic emotions (surprise, anger, disgust, fear, sadness and happiness) in a scale from 0 to 10 is also provided by the system. MixedEmotions tools are not available for Basque, but keeping in mind that the text to be said is known beforehand, the translation from Basque to English can be performed in advance.

As a result this module outputs a descriptor that combines the information obtained from these two tools. The descriptor is composed by the polarity label (negative, neutral or positive) extracted from the text by using EliXa and the VAD numbers (each one in [0,10] range) obtained with MixedEmotions tools.

### 3.2 How Is the Emotion Selected?

From the two pieces of information returned by the text sentiment analyzer we want to assess the emotion of a text in the sadness-happiness continuum, computing its numerical value. We have decided this numerical value to fall in the [0,10] scale, where 0 corresponds to the maximum sadness and 10 to the maximum happiness. The *Valence* value could roughly be translated to this scale, but it is not always the case that high values in valence as returned by MixedEmotions translate to positive polarity returned by EliXa. As we are specially interested in Basque language, and EliXa can handle it, we give more weight to EliXa's assessment.

The [0,10] emotion scale has been divided into three parts, where the interval [0,4.5] corresponds to negative polarity, [4.5,6.5] to neutral polarity and (6.5,10] to positive polarity. This division of the scale comes from the observation of the *Valence* values returned by MixedEmotions in a set of sentences. The aim is to directly translate the *Valence* into the sadness-neutral-happiness scale with a few caveats. Our approach is the following one: first we analyze the polarity of the text according to EliXa, and if the *Valence* according to MixedEmotions lies in the interval corresponding to the polarity, then the *Valence* value is used as it is. Otherwise, the limit value of the closest interval is chosen.

<sup>2</sup> <http://mixedemotions.insight-centre.org/>.

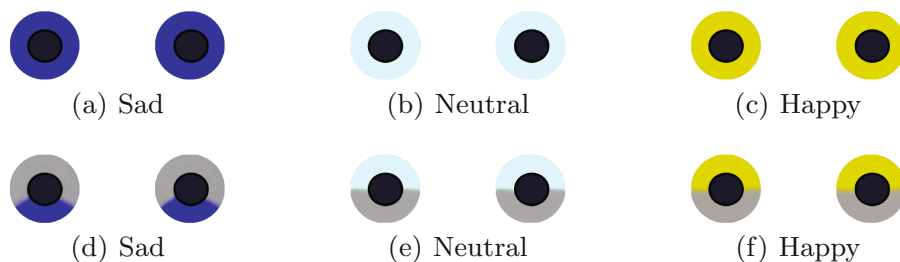
### 3.3 How Is the Emotion Displayed?

Once the analysis of the emotion is done, the translation from *Valence-Arousal-Dominance* space to a facial expression, voice intonation, and body expression must be done. Each way of expression has an associated module which makes that translation. Facial expression is displayed by the *Eyes Lightning Controller*, voice intonation variation is made by the speech synthesizer of the *Chatting* behavior, and body expression is showed by the *Gesture Controller*.

#### Eyes Lightning Controller

Inspired by human eyes, the eyes of NAO are composed by two rings of LEDs (8 LED per eye) with a black pupil inside. Each LED can be individually controlled for different color, intensity and duration. Johnson et al. [8] demonstrate that NAO’s eyes can be used to express emotions. The LED lighting configuration (color selection) used in our approach is based in the color-emotion study they carried out. Our contribution over their work is that in our approach the color and the intensity of the color changes depending on the *Valence* value of the emotion, and in their approach the emotion intensity is not taken into account. Sadness is displayed by a dark blue-greenish color that varies from RGB(0, 0, 255) to RGB(0, 255, 255), neutral is displayed by a light blue-white color from RGB(127, 255, 255) to RGB(255, 255, 255) and happiness is displayed by a yellow color from RGB(76, 76, 0) to RGB(255, 255, 0).

Moreover we have defined two sets of eye LED patterns in order to show emotions in two different ways. The *LED Pattern 1* makes use of the whole eye, whilst the *LED Pattern 2* uses only half of the eye. The second one was inspired by facial expressions displayed in cartoons (see Fig. 1). The experiments carried out (see Sect. 4) show us that the emotions are better understood through the second pattern.



**Fig. 1.** Eye LED patterns. *LED Pattern 1*: a, b, c. *LED Pattern 2*: d, e, f

#### Chatting Behavior

Previously we developed [18] a “Chatting” behavior that provides robots with the necessary skills for natural interaction with the interlocutor. It makes use of different tools that make robots able to speak in Basque, Spanish, and English. For Basque language, AhoTTS [19] text-to-speech tool is used, which

has its own speech synthesizer. For Spanish and English, NAO's TTS tool is used instead, which employs ACAPELA<sup>3</sup> speech synthesizer.

The influence of the voice intonation in emotional expression is clearly argued in [20]. Their study shows that some emotions, such as fear, happiness and anger, are portrayed in a higher speech rate and also at a higher pitch than emotions such as sadness. AhoTTS offers different types of voice intonation for Basque, one for each of six basic emotions and another one for neutral emotion. We have used the happiness, neutral and sadness intonations to portray the three emotions available in our system. Unfortunately NAO's speech synthesizer does not offer direct voice intonation selection as AhoTTS does. But it provides the option to setup some voice parameters such as pitch and speech rate, which can be tuned to obtain a different voice intonation than the standard provided. Pitch and speech rate parameter values have been experimentally defined for our system.

Our approach lies in selecting a different voice intonation relative to the emotion obtained after the emotion selection process. As mentioned before, till now our system only differentiates among sad, neutral and happy emotions so the voice intonation can only correspond to one of these three emotions.

### Gesture Controller

The way an emotion is reflected in the different parts of the body is well explained in [21]. For instance, the position of the head conveys sadness if it is tilted down, or happiness if it is tilted up. On the other hand, usually humans do not stay still while talking, we naturally gesticulate moving the hands. This action strengthens the message to be conveyed. According to Amaya et al. [22] the difference between the emotional and neutral movements lies in two parameters, speed and spatial amplitude.

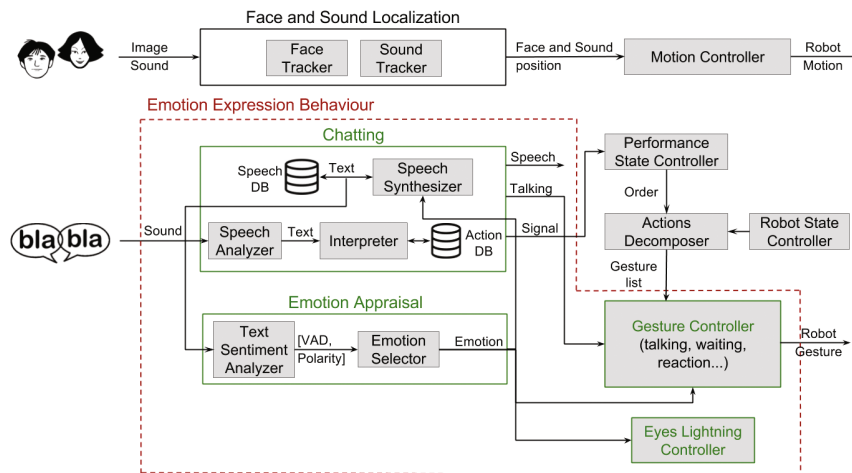
The robot body expression is managed by the *Gesture Controller*. It controls the head tilt, and the arm gestures to be reproduced (the set of the talking gestures used is the one proposed in [18]). For the head tilt of the robot, the *Valence* obtained is directly translated from VAD space to head pitch physical range ( $[0.35, -0.41]$  radian angle interval). However, for the arm gestures the approach is quite different. In a previous version of the system (see [18]), the sequence of the predefined gestures to be executed was randomly selected and the number of gestures was chosen according to the speech duration. Now the process is similar, but first the gestures are modified by increasing or decreasing the execution tempo in a 30% proportion of the *Valence*.

### 3.4 Global Control Architecture

The contribution of this work is a robotic emotion expression behavior, composed by several subsystems (*Emotion Appraisal*, *Eyes Lightning Controller*, *Chatting* and *Gesture Controller*) that make the robots being able to express emotions according to the sentiment of the speech, and make them more sociable. All

<sup>3</sup> <http://www.acapela-group.com/>.





**Fig. 2.** The global control architecture composed by several social behaviors

these subsystems have been implemented as ROS<sup>4</sup> modules. The aim of the system is to assess the emotional content of a written text and translate it to the body language and verbal expression of a humanoid robot. Figure 2 describes the architecture that brings together all the component behaviors.

## 4 Experimental Assessment

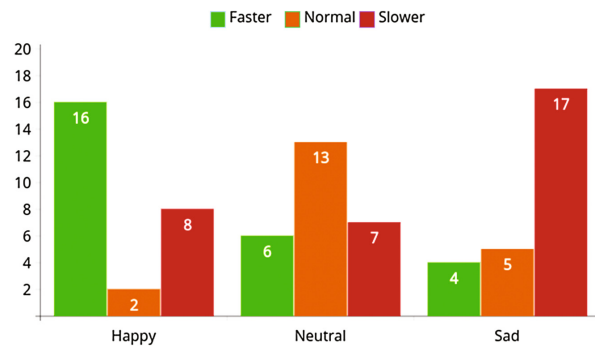
As previously stated, the robot is able to convey emotions through different parts of its body. An experiment has been designed to measure the meaningfulness of the eyes patterns and the impact of the arm movement. In order to evaluate these two features in a sounder manner we decided to leave the voice intonation aside. We considered that the voice intonation could disturb the perception of the relevant characteristics that are meant to be evaluated.

In this experiment people is faced to a NAO robot with different body language according to a predefined inner emotional state. This body language has been designed to express a happy, sad or neutral emotion. The goal of the experiment is to assess if the participants agree among themselves and with the researchers about the meaning of the robot body language.

As subjects of our experiment are 26 nine-years-old children attending a primary school class. It was hypothesized that the head position was the clearest indication of the robot emotional state and that the children would be able to recognize that a head looking up means the robot is happy and that the head looking down means the robot is sad. All the children recognized this as an obvious sign. Likewise, the LED lighting configuration as stated in the work described in [8] was also recognized in the same terms as desired. These surveys were made with all the children speaking at the same time, in an unstructured way to make the experiment more enjoyable to the children.

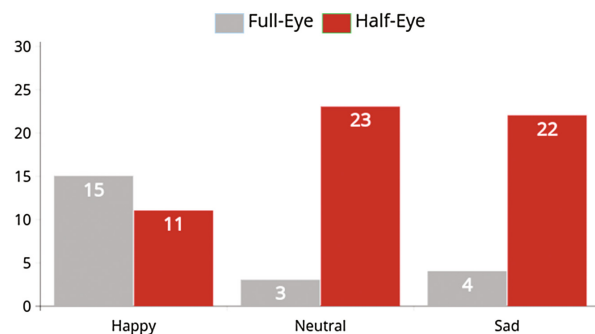
<sup>4</sup> <http://www.ros.org>.

Another two aspects of body language were evaluated through individual questionnaires: arm movement and eyes lighting. In the experiment involving arm movement the children were presented with the robot showing movement of its head meaning to be happy, neutral or sad, as previously recognized by the children. For each head position, the robot moved its arms with three different speeds, from faster to slower. We hypothesized that faster arm movement conveys happiness whilst slower movement expresses sadness. The children were asked to evaluate how they perceived the different speeds to convey a happy, neutral or sad emotion. Figure 3 shows the result of this experiment. As it can be seen, the children mainly agree with the supposition that faster arm movements correspond to a happier emotional state.



**Fig. 3.** Arm movement evaluation

In the experiment involving eye LED lighting, the goal is to assess if the children appreciate some difference in the strength of an emotion depending on whether all the LEDs in the eyes are switched on or only half of them. Figure 4 shows how the children see that activation of half of the eye is best suited to neutral and sad emotions, although happiness is best shown with all the LEDs switched on.



**Fig. 4.** Eye LED evaluation

## 5 Conclusions and Further Work

The emotional behavior presented in this work endows the robots with the ability to adapt their way to express different emotions according to the sentiment of the speech, and it makes the robots interact in a more social way in a human-robot environment. The approach of the system consists of assessing the emotional content of a written text and translate this content to the body language and verbal expression of a humanoid robot in the sadness-happiness continuum. The adaptive emotional system that we have developed combines different subsystems that makes a humanoid robot able to express emotions by adapting the head pitch, arm movement tempo of predefined gestures, eye LED lighting and voice intonation. The system is available in English, Spanish and Basque languages.

The body language implemented in the robot has been recognized as intended by a group of nine-years-old children. The children mainly agree with the supposition that faster arm movements correspond to a happier emotional state, and slower movements to a sadder emotional state. Regarding to the eyes lighting, questionnaires results show that the emotions are better understood through partial eyes lighting, except in the case of happiness.

There are several research lines that we will explore in future work. Regarding to the body expression, we will investigate the effect of moving the different parts of the body in the same way we have done with the head. It can be analyzed how the trunk and arms inclination affects in the emotion displayed. Furthermore, we have in mind to add to the system the possibility of showing more emotions, such as surprise, anger, disgust and fear. Experiments involving more children and adults are planned, in order to draw statistically sound results.

**Acknowledgment.** This work has been partially supported by the Basque Government (IT900-16), the Spanish Ministry of Economy and Competitiveness MINECO/FEDER (TIN 2015-64395-R, MINECO/FEDER,EU). Author Rodriguez has received the UPV/EHU research Grant PIF13/104.

## References

1. Knight, H.: Eight lessons learned about non-verbal interactions through robot theater. In: *Social Robotics*, pp. 42–51 (2011)
2. Breazeal, C., Siegel, M., Berlin, M., Gray, J., Grupen, R., Deegan, P., Weber, J., Narendran, K., McBean, J.: Mobile, dexterous, social robots for mobile manipulation and human-robot interaction. p. 27. *ACM* (2008)
3. Breazeal, C.: Emotion and sociable humanoid robots. *Int. J. Hum. Comput. Stud.* **59**(1–2), 119–155 (2003)
4. Cassell, J., Vilhjálmsón, H.H., Bickmore, T.: BEAT: the behavior expression animation toolkit. In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 477–486. *ACM* (2001)
5. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) *IWA 2006*. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006). doi:[10.1007/11821830\\_20](https://doi.org/10.1007/11821830_20)

6. Lhommet, M., Xu, Y., Marsella, S.: Cerebella: automatic generation of nonverbal behavior for virtual humans. In: AAAI, pp. 4303–4304 (2015)
7. Breazeal, C.: Designing sociable robots. In: Intelligent Robotics and Autonomous Agents. MIT Press, Cambridge MA, USA (2004)
8. Johnson, D.O., Cuijpers, R.H., van der Pol, D.: Imitating human emotions with artificial facial expressions. *Int. J. Soc. Robot.* **5**(4), 503–513 (2013)
9. Paradedda, R.B., Hashemian, M., Rodrigues, R.A., Paiva, A.: How facial expressions and small talk may influence trust in a robot. In: Agah, A., Cabibihan, J.-J., Howard, A.M., Salichs, M.A., He, H. (eds.) ICSR 2016. LNCS (LNAI), vol. 9979, pp. 169–178. Springer, Cham (2016). doi:[10.1007/978-3-319-47437-3\\_17](https://doi.org/10.1007/978-3-319-47437-3_17)
10. Feldmaier, J., Marmat, T., Kuhn, J., Diepold, K.: Evaluation of a RGB-LED-based emotion display for affective agents. arXiv preprint [arXiv:1612.07303](https://arxiv.org/abs/1612.07303) (2016)
11. Beck, A., Cañamero, L., Bard, K.A.: Towards an affect space for robots to display emotional body language. In: IEEE RO-MAN, pp. 464–469 (2010)
12. Beck, A., Cañamero, L., Hiolle, A., Damiano, L., Cosi, P., Tesser, F., Sommovilla, G.: Interpretation of emotional body language displayed by a humanoid robot: a case study with children. *Int. J. Soc. Robot.* **5**(3), 325–334 (2013)
13. Tielman, M., Neerincx, M., Meyer, J.J., Looije, R.: Adaptive emotional expression in robot-child interaction. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, pp. 407–414. ACM (2014)
14. McColl, D., Nejat, G.: Recognizing emotional body language displayed by a human-like social robot. *Int. J. Soc. Robot.* **6**(2), 261–280 (2014)
15. Bretan, M., Hoffman, G., Weinberg, G.: Emotionally expressive dynamic physical behaviors in robots. *Int. J. Hum. Comput. Stud.* **78**, 1–16 (2015)
16. Ekman, P.: Are there basic emotions? *Psychol. Rev.* **99**(3), 550–553 (1992)
17. San Vicente, I., Saralegi, X., Agerri, R., Sebastián, D.S.: EliXa: a modular and flexible ABSA platform. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), pp. 748–752 (2015)
18. Rodriguez, I., Astigarraga, A., Ruiz, T., Lazkano, E.: Singing minstrel robots, a means for improving social behaviors. In: International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, pp. 2901–2907, May 2016
19. Hernaez, I., Navas, E., Murugarren, J., Etxebarria, B.: Description of the AhoTTS system for the Basque language. In: 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, Perthshire, Scotland (2001)
20. Bänziger, T., Scherer, K.R.: The role of intonation in emotional expressions. *Speech Commun.* **46**(3), 252–267 (2005)
21. Lhommet, M., Marsella, S.C.: Expressing emotion through posture and gesture. In: The Oxford Handbook of Affective Computing, pp. 273–285. Oxford University Press (2015)
22. Amaya, K., Bruderlin, A., Calvert, T.: Emotion from motion. In: Graphics Interface. vol. 96, Toronto, Canada, pp. 222–229 (1996)

## 12.7 Spontaneous Talking Gestures Using Generative Adversarial Networks

**Title:** Spontaneous Talking Gestures Using Generative Adversarial Networks

**Authors:** I. Rodriguez, J. M. Martínez-Otzeta, I. Irigoien, E. Lazkano

**Journal:** Robotics and Autonomous Systems

**Publisher:** Elsevier

**DOI:** 10.1016/j.bica.2018.07.014

**Year:** 2018

# Spontaneous Talking Gestures Using Generative Adversarial Networks

Igor Rodriguez<sup>a,\*</sup>, José María Martínez-Otzeta<sup>a</sup>, Itziar Irigoien<sup>a</sup>, Elena Lazkano<sup>a</sup>

<sup>a</sup>*Computer Science and Artificial Intelligence  
Faculty of Informatics (UPV/EHU)  
Manuel Lardizabal 1, 20018 Donostia*

---

## Abstract

This paper presents a talking gesture generation system based on Generative Adversarial Networks, along with an evaluation of its adequateness. The talking gesture generation system produces a sequence of joint positions of the robot's upper body which keeps in step with an uttered sentence. The suitability of the approach is demonstrated with a real robot. Besides, the motion generation method is compared with other (non-deep) generative approaches. A two-step comparison is made. On the one hand, a statistical analysis is performed over movements generated by each approach by means of Principal Coordinate Analysis. On the other hand, the robot motion adequateness is measured by calculating the end effectors' jerk, path lengths and 3D space coverage.

*Keywords:* Social robotics, generative learning models, motion generation, principal coordinate analysis, body language expression, generative adversarial networks

---

## 1. Introduction

Social robotics [1] aims to provide robots with artificial social intelligence to improve human-machine interaction and to introduce them in complex human contexts. The demand for sophisticated robot behaviors requires to model

---

\*Corresponding author

*Email address:* [igor.rodriiguez@ehu.eus](mailto:igor.rodriiguez@ehu.eus) (Igor Rodriguez)

*URL:* [www.sc.ehu.eus/ccwrobot](http://www.sc.ehu.eus/ccwrobot) (Igor Rodriguez)

5 and implement human-like capabilities to sense, to process, and to act/interact naturally by taking into account emotions, intentions, motivations, and other related cognitive functions.

Naturally, speech plays a relevant role to convey emotions, and human voice can be shaped in very complex ways. Some works show [2] that the level of  
10 trust of the human with respect to the robot is higher when the robot's gaze is in the direction of the interlocutor. Besides, talking involves spontaneous gesticulation; postures and movements are relevant for social interactions even if they are subjective and culture dependent. Gestures (head, hands and arms movements) are used both to reinforce the meaning of the words and to express  
15 feelings through non-verbal signs. The impressive realistic character animations developed within the fields of computer graphics and virtual agents reflect the importance of synchronizing synthesized speech with non-verbal behavior [3][4].

No doubt, body language can disclose as much information as words. And a high expression capability can be achieved with a low number of degrees of  
20 freedom (DoF). For instance, Anki's Cozmo [5] a tiny robot designed to interact with by playing, shows an impressive body expression. A kind of shovel that it uses for manipulation purposes adds arm-level expression in a wheeled platform. Shimi [6], a smart-phone enabled robotic musical companion far from human morphology, expresses emotion rather differently, using a faceless body. But the  
25 advent of anthropomorphic robots has launched researchers to investigate and develop human-like body language expression in robots. Building natural robot behaviors enhances the expressiveness of robots and improves sociability.

L'hommet and Marsella [7] discuss body expression in terms of postures, movements and gestures. Gestures, defined as movements that convey informa-  
30 tion intentionally or not, are categorized as emblems, illustrators and adaptors. Emblems are gestures deliberately performed by the speaker that convey meaning by themselves and are again culture dependent. Illustrators are gestures accompanying speech, that may (pointing to an object) or may not (beats) be related to the semantics of the speech. Lastly, adaptors or manipulators belong  
35 to the gesture class that does not aid in understanding what is being said, such

as tics or restless movements.

The goal of this paper is to develop a talking gesticulation behavior that resembles humans in terms of naturalness and expressiveness. Within the work context of this paper, no semantic meaning is being extracted from the spoken  
40 text and, thus, the gesture repertoire of the robot is limited to body adaptors applied to talking behavior.

In [8] the authors randomly selected gestures from a set of movements previously compiled, but that approach is prone to produce repetitive movements and can result in unnatural jerky expression. A generative approach should  
45 allow to create novel movements while retaining their nature, overcoming the limitations of a predefined set of gestures. In conclusion, a generative approach should allow for a more spontaneous behavior.

Generative Adversarial Networks (GANs) [9] are deep generative models capable to implicitly acquire the probability density function in the training  
50 data, as deep learning methods are able to automatically discover the internal structure of datasets, by learning multiple levels of abstraction [10]. They can overcome some problems associated with other generative models, as the need for explicitly searching in the space of a given family of probability distributions, like for instance in Gaussian Mixture Models. GANs are also subjectively  
55 regarded as producing better samples than other methods [11], and therefore are chosen for the task at hands.

The rest of the paper is structured as follows: Section 2 summarizes the related work found in the literature. Next, Section 3 describes how GANs work. In Section 4 the experimental setup is described, while the obtained results from  
60 the point of view of the robot's behavior are presented in Section 5. The performance analysis of the applied method is shown in Section 6. There, the motion generation method is compared with other (non-deep) generative approaches. A two step comparison is made. On the one hand, a statistical analysis is performed over movements generated by each approach by means of Principal  
65 Coordinate Analysis. On the other hand, the robot motion adequateness is measured by calculating the end effectors' jerk, path lengths and the 3D space



coverage. Finally, conclusions are presented in Section 7.

## 2. Related Work

Generative models are probabilistic models capable of generating all the values for a phenomenon. Unlike discriminative models, they are able to generate not only the target variables but also the observable ones [12]. They are used in machine learning to (implicitly or explicitly) learn the distribution of the data for generating new samples. There are many types of generative models. Bayesian Networks (BNs) [13], Gaussian Mixture Models (GMMs) [14] and Hidden Markov Models (HMMs) [15] are well known probability density estimators. Deep learning techniques have been applied to generative models, giving rise to deep generative models. A taxonomy of such models can be found in [11].

Applications of generative models range from photo-realistic single image super-resolution [16] and text-to-image synthesis [17] to handwriting sequences generation [18] using recurrent neural networks (RNN) or speech synthesis [19] based on WaveNet [20], an autoregressive deep generative model. In [21] the authors propose Deep Generative Spacial Models (DGSM), the first application of Sum-Product Networks to the domain of robotics. A generative model is able to learn a single, universal model of the robot’s spatial environment. In astronomy GANs are becoming popular for improving images [22].

Generative models are also being used for motion generation. In [23] the authors propose the combination of Principal Component Analysis (PCA) [24] and HMMs for encoding different movement primitives to generate humanoid motion. A. K. Tanwani [12] uses HSMM (Hidden Semi-Markov Models) for learning robot manipulation skills from humans. Focusing on social robotics, some generative approaches are being applied with different means. In [25] Manfrè et al. use HMMs for dance creation and in a later work they try variational auto-encoders again for the same purpose [26]. Regarding the use of adversarial networks, Gupta et al. [27] extend the use of GANs to generate socially acceptable motion trajectories in crowded scenes in the scope of self-

driving cars.

### 3. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [9] are semi-supervised emerging models that basically learn how to generate synthetic data from the given training data. A GAN network is composed by two different interconnected networks. The *Generator* ( $G$ ) network generates possible candidates so that they are as similar as possible to the training set. The second network, known as *Discriminator* ( $D$ ), judges the output of the first network to discriminate whether its input data are “real”, namely equal to the input data set, or if they are “fake”, that is generated to trick with false data. The general architecture of a GAN with the  $G$  and the  $D$  networks is shown in Figure 1. The generator is typically a deconvolutional neural network, while the discriminator is a convolutional one.

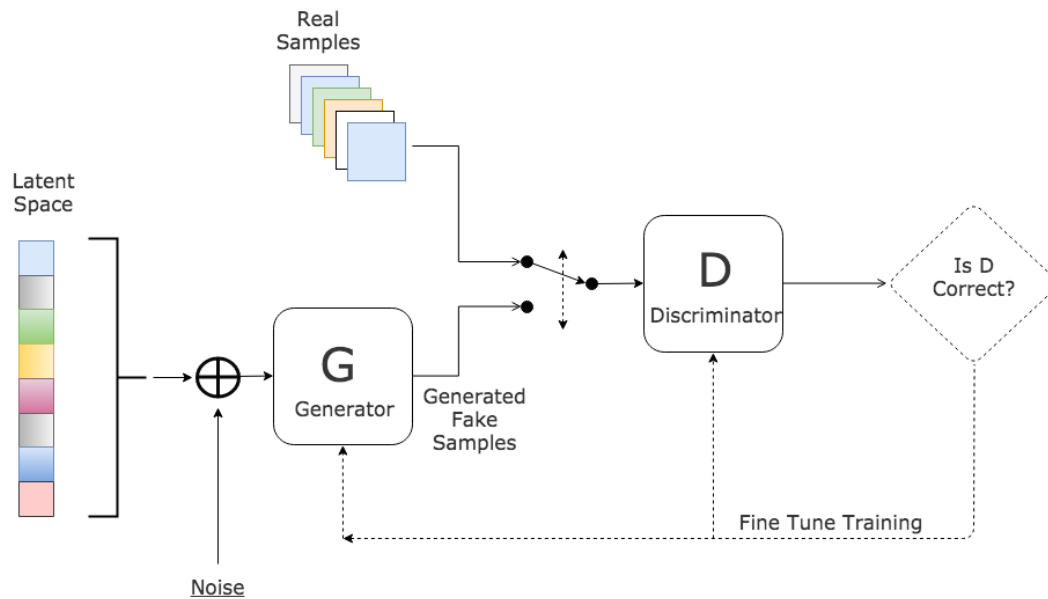


Figure 1: Description of GAN architecture

In the first step,  $D$  takes as input both, real data and fake data, and returns for each sample its probability to come from real data. In the second step, the  $G$  network is trained. While the parameters of  $D$  are fixed, in each epoch, the

weights of the  $G$  network are updated to let the discriminator results on the sample generated by  $G$  as near as possible to 1. That is, this second step is aimed to modify the  $G$  network in order to be able to generate samples that can trick the  $D$  network.

115 The  $G$  network is never exposed to real data, the only manner to enhance its generation capability is through the interaction with  $D$  by means of the output. Instead,  $D$  has access to both, real data and fake data, and produces as output the ground truth to know if the data came from the generator or the dataset. The discriminator's output value is exploited by the generator to enhance the  
120 quality of the forgery data.

Back-propagation is applied in both networks to enhance the accuracy of the generator to produce valid movements; on the other side, the discriminator becomes more skilled to flag false data.

## 4. Experimental Setup for Talking Gesture Creation

### 125 4.1. Robotic Platform and Framework

The robotic platform employed in the performed experiments is a Pepper robot developed by Softbank Robotics <sup>1</sup>. Pepper is a human-like torso that it is fitted onto a holonomous wheeled platform. It is equipped with full-color RGB LEDs (placed in eyes, ears and shoulders), three cameras, and several sensors  
130 located in different parts of its body that allow for perceiving the surrounding environment with high precision. It heights 120 cm and has 20 DoFs.

NAOqi <sup>2</sup> is the name of the SDK (Software Development Kit) provided by Aldebaran, that runs on the robot and controls it. NAOqi is designed so that modules can be run independently across multiple machines and robots. Each  
135 module has an external API (i.e., functions and parameters) that other modules can call. Currently, our robot is controlled using the *naoqi\_driver* driver that

---

<sup>1</sup><https://www.aldebaran.com/en/robots/pepper>

<sup>2</sup><http://doc.aldebaran.com/2-5/naoqi/index.html>

wraps the needed parts of NAOqi API and makes it available in ROS <sup>3</sup>.

#### 4.2. Talking Movement Definition

The fixed picture of the position and orientation of the joints of the robot is called pose. In our work, it is composed by 14 float numbers. A pose is thus represented by a set of 14 joint values, comprising robot's head, hands and arms (Table 1). Pelvis, knee and wheeled base information were ignored because those elements are not involved in talking adaptors. As we are interested in generating movements, i.e., a sequence of poses, the input to the learning process to any generative model has to take into account the temporal sequence of poses. We have defined the unit of movement as a sequence of four consecutive poses. This unit of movement, a vector of 56 float numbers (14 from each pose and four poses concatenated) will be the desired output of the generative model (see Table 2). The training set of such model is thus a set of units of movement. This training set is generated taking four consecutive poses from a database of poses, ordered according to their appearance in a movement. For example, if a movement consists of the following temporally consecutive poses  $(P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8)$ , there are five possible units of movement:  $(P_1, P_2, P_3, P_4)$ ,  $(P_2, P_3, P_4, P_5)$ ,  $(P_3, P_4, P_5, P_6)$ ,  $(P_4, P_5, P_6, P_7)$  and  $(P_5, P_6, P_7, P_8)$ . Our database consists of non-overlapping units of movement, so only  $(P_1, P_2, P_3, P_4)$  and  $(P_5, P_6, P_7, P_8)$  would be taken into account.

At the end, when a whole movement is required, in order to be executed by the robot, a number of units of movement are asked to the generative model, one after another. How many of these units of movement form a whole movement will depend on the desired temporal length of the movement, which, at its time, will depend of the temporal length of the audio intended to be pronounced by the robot.

---

<sup>3</sup>[www.ros.org](http://www.ros.org)

Head	$H_\gamma(J_1)$	$H_\beta(J_2)$
Right Shoulder	$RS_\beta(J_3)$	$RS_\alpha(J_4)$
Right Elbow	$RE_\beta(J_5)$	$RE_\gamma(J_6)$
Right Wrist	$RW_\gamma(J_7)$	
Right Hand	$RH(J_8)$	
Left Shoulder	$LS_\beta(J_9)$	$LS_\alpha(J_{10})$
Left Elbow	$LE_\beta(J_{11})$	$LE_\gamma(J_{12})$
Left Wrist	$LW_\gamma(J_{13})$	
Left Hand	$LH(J_{14})$	

Table 1: Format of a robot pose.  $\alpha$ : roll,  $\beta$ : pitch,  $\gamma$ : yaw angles. Hands can be opened or closed. A movement is thus composed of 4 consecutive poses. In parenthesis, a more convenient notation for formal use.

#### 4.3. Training Database

The training dataset given to the  $D$  network to learn the distribution space  
of the data is created from gestures obtained from the default animations of the  
165 NAOqi API. We have chosen a subset of those gestures that can be used for  
accompanying the speech, and can be performed individually or in composition  
to constitute complex sequences of gestures.

In order to collect data, we sampled the poses during the selected NAOqi  
170 animations with a frequency of 4 Hz. About 1200 units of movement were  
collected for training.

$$J_1(t) \cdots J_{14}(t), J_1(t + \Delta t) \cdots J_{14}(t + \Delta t), J_1(t + 2\Delta t) \cdots J_{14}(t + 2\Delta t), J_1(t + 3\Delta t) \cdots J_{14}(t + 3\Delta t)$$

Table 2: Characterization of a unit of movement.  $\Delta t$  depends of the data sampling frequency

#### 4.4. GAN Setup

The discriminator network is thus trained using the previously mentioned  
data to learn the distribution space of the data. On the other hand, the gener-

175 ator is seeded through a random input with a uniform distribution in the range  
[−1, 1] and with a dimension of 100. The *Generator* intends to produce as out-  
put gestures that belong to the real data distribution and that the *Discriminator*  
network would not be able to correctly pick out as generated.

180 Regarding GAN’s hyper-parameters, after several experiments, we setup a  
batch size of 16, a learning rate of 0.0002, Adam [28] as the optimization method,  
and  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$  as its parameters. We trained the network for  
2000 epochs.

## 5. Results

The obtained robot performance is shown in the following two videos:

- 185
1. Video<sup>4</sup> shows the evolution of the robot behavior during different steps  
of the training process. The final number of epochs has been empirically  
defined, observing the behavior of the robot.
  2. Video<sup>5</sup> qualitatively demonstrates the adequateness of the approach by  
showing how the robot behaves while talking.

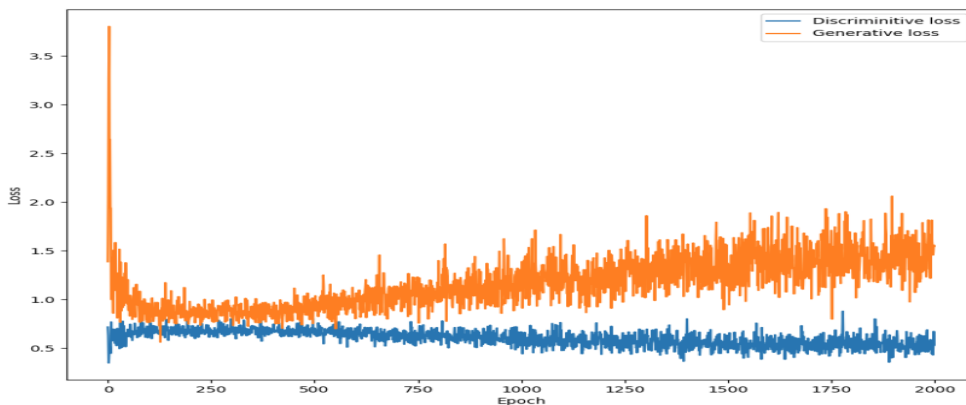


Figure 2: Loss functions evolution over 2000 epochs

---

<sup>4</sup><https://www.youtube.com/watch?v=AW3BmfS7DIY>

<sup>5</sup><https://www.youtube.com/watch?v=KVyTbFEMcHE>

## 190 6. Performance Analysis

In order to quantitatively measure the appropriateness of the obtained motion generation system, it has been compared with other generative approaches have been used:

1. Randomly generated movements (RND): the simplest way of generating  
195 new movements is to concatenate random poses from the already existing set.
2. Gaussian Mixture Models: those attempt to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset.
- 200 3. Hidden Markov Models: HMMs can be considered a generalization of mixture models where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.

The random movements have been produced choosing a random pose from the  
205 already existing set and concatenating it with the three following ones. Referring to the previously mentioned possible movement  $(P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8)$ , a possible random unit of movement not present in the training set would be  $(P_3, P_4, P_5, P_6)$ . The GMM and HMM models were learned using the same database utilized for training the GAN. Bayesian Information Criterion (BIC)  
210 [29] optimization was used for GMM model selection and the best model used tied covariance matrix (the same covariance matrix is shared by all Gaussians) and 24 components. Regarding the HMM, the chosen model was a Gaussian HMM with 10 hidden units and full covariance matrix.

A two step comparison has been made along the different approaches: Prin-  
215 cipal Coordinate Analysis and robot motion analysis.

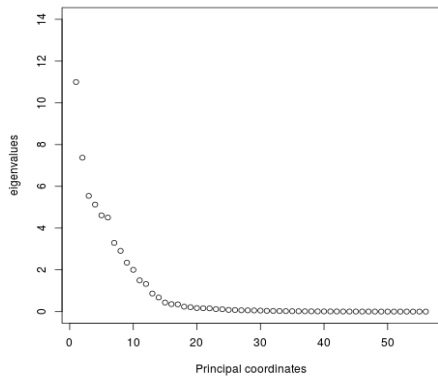
### *6.1. Principal Coordinate Analysis*

1000 units of movement were sampled for the different models. This gives a  $1000 \times 56$  data matrix for each method where columns represent the positions of

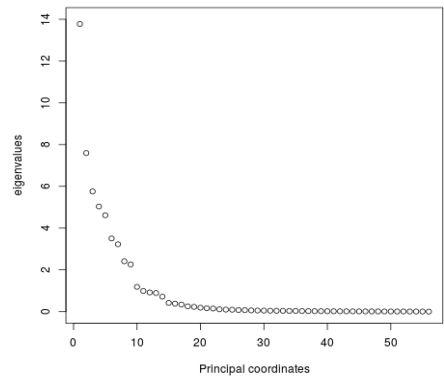
the joints along the unit of movement ( $J_i(t + k\Delta t), i = 1, \dots, 14, k = 0, \dots, 3$ )  
 (see Tables 1 and 2). The structure underlying the units of movement was  
 analyzed considering the relationship between the joints. First, correlation dis-  
 tances [30] between joints for the NAOqi original units of movement, RND,  
 GAN, GMM, and HMM were calculated. In order to get comparable results  
 the distance matrices were scaled so that their geometric variability were equal  
 to 1. Then, a Principal Coordinates Analysis [31] was carried out on each dis-  
 tance matrix. In Figure 3 the variability decompositions corresponding to the  
 distance matrices are shown. This decomposition shows the essence of the struc-  
 ture that lies between joint poses for each type of motion generation method.  
 For instance, the joints behavior for NAOqi and GAN units of movement are  
 embedded, broadly, in a 15-dimensional euclidean space where the first two  
 contain remarkably the most variability. For HMM units of movement too, the  
 first two dimensions are remarkably the most important ones but more than 40  
 dimensions are needed to embed the joints behavior. For RND and GMM units  
 of movement, the underlying structure defined by the joints are very different  
 from the aforementioned ones.

The representations of the joints on the corresponding two axes also show  
 that there are different underlying structures of the joints for the different type  
 of units of movement generations (see Figure 4). For joints related the original  
 NAOqi units of movement a smoothness in the sequence of poses is appreciated  
 since the sequence of joint units of movement  $J_i(t), \dots, J_i(t + 3\Delta t)$  are clustered  
 tightly, for all joints 1 to 14. For the GAN units of movement, a similar pattern  
 of smoothness is present but the clustering of sequential poses is more loose.  
 On the contrary, for the rest type of units of movement this smoothness pattern  
 is lost. Moreover, it is interesting to notice that in general head joints poses  
 are located in the center of the plane and that for RND and GMM generated  
 movement there is a an overall ordering of the ‘Left’ and ‘Right’ joints poses on  
 the plane.

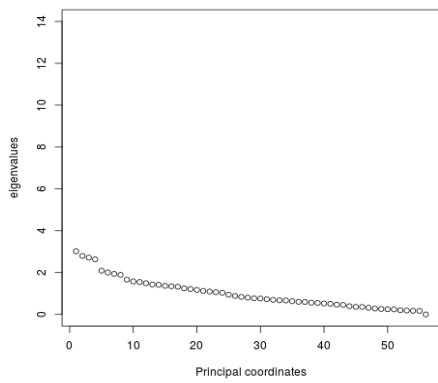




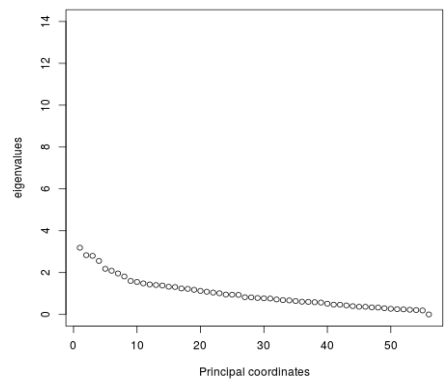
(a) Original database



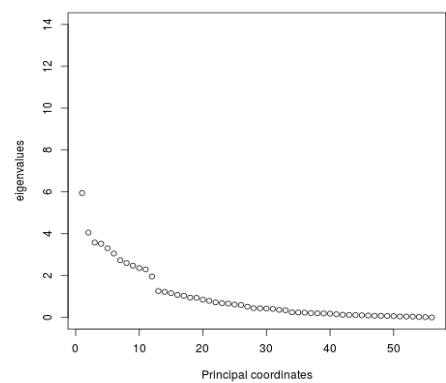
(b) GAN



(c) Random



(d) GMM



(e) HMM

Figure 3: Variability decomposition after applying Principal Coordinate Analysis on the correlation distance matrices



## 6.2. Robot Motion Analysis

For quantitatively measuring the quality of the generated movements three  
250 measures are presented:

- Norm of Jerk: as mentioned in the introduction, the goal is to generate spontaneous smooth movements. The norm of Jerk is a smoothness measure based on root mean square (RMS) jerk quantification [32]. It is calculated according to Equation 1.

$$jerk = \frac{1}{T} \sum_{t=1}^T \|\dot{accel}_t\| \quad (1)$$

- 255
- 3D space coverage: small *jerk* values capture the absence of sudden acceleration changes and are thus, desirable. But it must also be taken into account the surface reached by the end effectors. The more 3D space covered, the more different the generated patterns are allowed to be. A natural measure would be the volume of the convex hull defined by the positions in time, but as only the points in the surface of the hull take part in the volume computation, another measure taking into account also the  
260 inner points would be more convenient. We have chosen the dispersion measure given by the mean distance of the points to their centroid, and have called it as *disp*. It is calculated according to Equation 2.

$$disp = \frac{1}{T} \sum_{t=1}^T \|\bar{x}_t - \frac{1}{T} \sum_{k=1}^T \bar{x}_k\| \quad (2)$$

- 265
- Length of the generated paths: the length of the path (*lpath*) described by the positions of the hands during time is also another interesting measure. Lower *jerk* values would lead to lower *lpath* values, as the movements would be smoother. The measure (*lpath*) is computed as Equation 3.

$$lpath = \sum_{t=2}^T \|\bar{x}_t - \bar{x}_{t-1}\| \quad (3)$$

In order to obtain those measures, 10 sentences have been selected (of approximately 25 seconds of duration) and executed, and the 3D coordinates of the end effectors (i.e left and right hands) with respect to the pelvis have been calculated while talking. Results are shown in Table 3. GAN trajectories have smaller jerk values, together with smaller path lengths but, at the same time, the corresponding dispersion measure value is the highest for both hands.

		RND	<b>GAN</b>	GMM	HMM
Lhand	$E_{jerk}$	0.022	0.017	0.027	0.030
	$\sigma_{jerk}$	0.002	0.003	0.003	0.002
	$E_{disp}$	0.066	0.080	0.059	0.058
	$\sigma_{disp}$	0.010	0.018	0.005	0.008
	$E_{lpath}$	2.079	1.5755	2.136	2.2598
	$\sigma_{lpath}$	0.2443	0.1776	0.2358	0.149
Rhand	$E_{jerk}$	0.030	0.026	0.056	0.046
	$\sigma_{jerk}$	0.003	0.007	0.010	0.006
	$E_{disp}$	0.083	0.086	0.082	0.074
	$\sigma_{disp}$	0.009	0.016	0.011	0.006
	$E_{lpath}$	3.1742	2.058	4.245	3.6974
	$\sigma_{lpath}$	0.292	0.4744	0.549	0.3509

Table 3: Mean and deviation values for each measure

Figure 5 reflects these results. GAN and HMM paths of both hands for one talking session are reproduced, together with that of the original NAOqi’s ones. Clearly, GAN produces trajectories more similar in shape to the non generative (and thus, repetitive) original gesture set.

Note that left and right hand movements are quite different. This characteristic, inherent in the recorded training database (NAOqi), is inherited by the generative methods.

## 7. Conclusions and further work

In this work a talking gesture generation system has been developed using GANs. The suitability of the approach is demonstrated with the real robot. Moreover, in order to quantitatively measure the goodness of the method, it has been compared with other (non-deep) generative approaches in terms of Principal Coordinate Analysis and robot motion analysis. Results suggest that GANs are a suitable method for generating robot movements that capture the essence of the predefined API gestures, while allowing more variability and, overall, giving a subjective impression of naturalness. In this research this impression is reinforced by the quantitative measures aforementioned.

As further work, we intend to substitute the NAOqi animation set used as training database by a richer gesture set by recording talking humans with a 3D camera. The approach can also be extended to other kind of gestures.

## Acknowledgements

This work has been partially supported by the Basque Government (IT900-16), the Spanish Ministry of Economy and Competitiveness MINECO/FEDER (TIN 2015-64395-R, MINECO/FEDER, EU). Author I. Rodriguez has received the UPV/EHU research Grant PIF13/104.

## References

- [1] C. Breazeal, *Designing sociable robots*, Intelligent Robotics and Autonomous Agents, MIT Press, Cambridge MA, USA, 2004.
- [2] R. B. Paradedá, M. Hashemian, R. A. Rodrigues, A. Paiva, How facial expressions and small talk may influence trust in a robot, in: *International Conference on Social Robotics*, Springer, 2016, pp. 169–178.
- [3] M. Neff, M. Kipp, I. Albrecht, H.-P. Seidel, Gesture modeling and animation based on a probabilistic re-creation of speaker style, *ACM Trans. Graph.* 27 (1) (2008) 5:1–5:24.

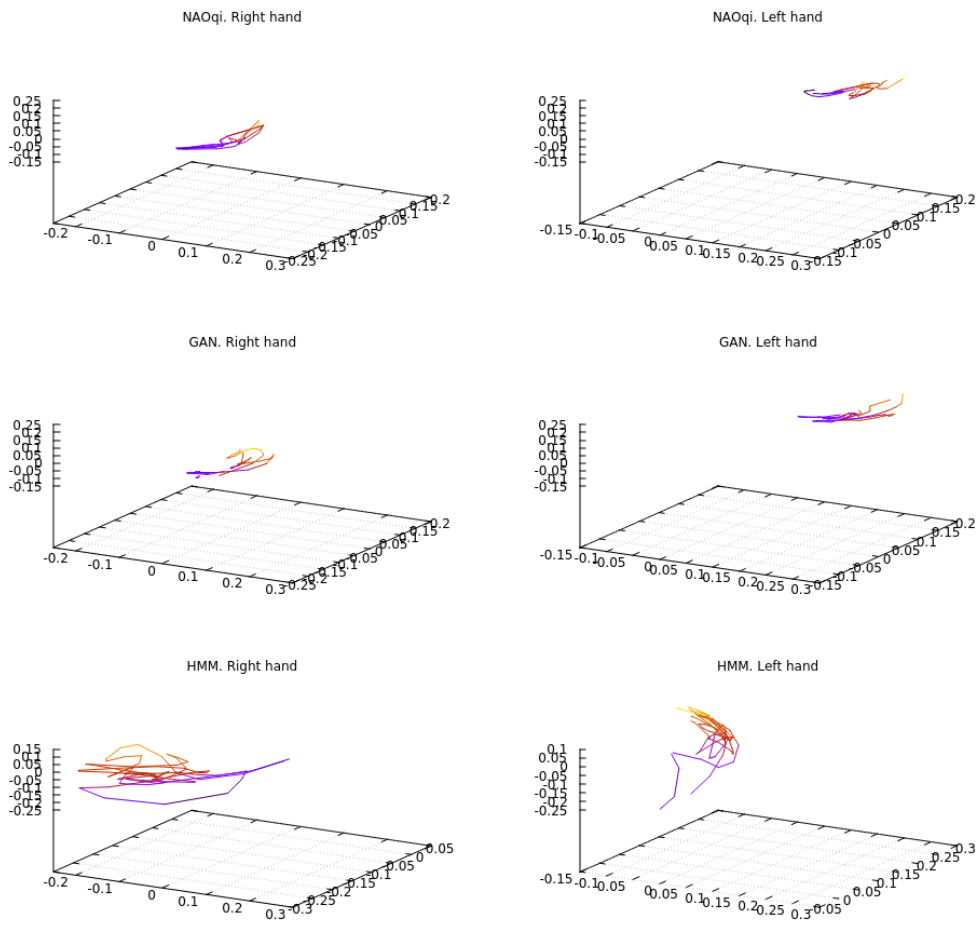


Figure 5: End effectors motion paths for one run

- [4] J. Cassell, H. H. Vilhjálmsón, T. 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.
- [5] Anki, Cozmo, <http://www.anki.com/en-us/cozmo/cozmo-tech> (accessed January 24, 2017).
- [6] M. Bretan, G. Hoffman, G. Weinberg, Emotionally expressive dynamic physical behaviors in robots, *International Journal of Human-Computer Studies* 78 (2015) 1–16.
- [7] M. Lhommet, S. Marsella, *The Oxford Handbook of Affective Computing*, Oxford University Press, 2015, Ch. Expressing Emotion Through Posture and Gesture, pp. 273–285.
- [8] I. Rodriguez, A. Astigarraga, T. Ruiz, E. Lazkano, Singing minstrel robots, a means for improving social behaviors, in: *IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2902–2907.
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [10] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
- [11] I. Goodfellow, NIPS Tutorial: Generative Adversarial Networks, ArXiv e-prints.
- [12] A. K. Tanwani, Generative models for learning robot manipulation, Ph.D. thesis, *École Polytechnique Fédéral de Laussane (EPFL)* (2018).
- [13] J. M. G. Enrique Castillo, A. S. Hadi, *Learning Bayesian Networks. Expert Systems and Probabilistic Network Models*, Monographs in computer science. New York: Springer-Verlag, 1997.
- [14] B. Everitt, D. Hand, *Finite mixture distributions*, Chapman and Hall, 1981.

- [15] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, in: Proceedings of the IEEE, Vol. 77, 1989, pp. 257–286.
- [16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, arXiv preprint.
- [17] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: M. F. Balcan, K. Q. Weinberger (Eds.), Proceedings of The 33rd International Conference on Machine Learning, Vol. 48 of Proceedings of Machine Learning Research, PMLR, New York, New York, USA, 2016, pp. 1060–1069.
- [18] A. Graves, Generating sequences with recurrent neural networks, Tech. rep., Cornell University (2013).
- [19] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. C. Rus, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, D. Hassabis, Parallel wavenet: Fast high-fidelity speech synthesis, Tech. rep., Google Deepmind (2017).  
URL <https://arxiv.org/abs/1711.10433>
- [20] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499.
- [21] A. Pronobis, R. P. N. Rao, Learning deep generative spatial models for mobile robots, Tech. rep., Cornell University (2017).
- [22] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, G. K. Santhanam, Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit, Monthly Notices



- of the Royal Astronomical Society: Letters 467 (1) (2017) L110–  
365 L114. [arXiv:/oup/backfile/content\\_public/journal/mnrasl/467/1/10.1093\\_mnrasl\\_slx008/1/slx008.pdf](https://arxiv.org/abs/10.1093/mnrasl_slx008/1/slx008.pdf), doi:10.1093/mnrasl/slx008.  
URL <http://dx.doi.org/10.1093/mnrasl/slx008>
- [23] J. Kwon, F. C. Park, Using hidden markov models to generate natural hu-  
manoid movement, in: IEEE/RSJ International Conference on Intelligent  
370 Robots and Systems, 2006.
- [24] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemo-  
metrics and intelligent laboratory systems 2 (1-3) (1987) 37–52.
- [25] A. Manfrè, I. Infantino, F. Vella, S. Gaglio, An automatic system for hu-  
manoid dance creation, Biologically Inspired Cognitive Architectures 15  
375 (2016) 1–9.
- [26] A. Augello, E. Cipolla, I. Infantino, A. Manfrè, G. Pilato, F. Vella, Creative  
robot dance with variational encoder, CoRR abs/1707.01489.
- [27] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, A. Alahi, Social GAM: so-  
cially accpetable trajectories with generaive adversarial networks, in: Com-  
puter Vision and Pattern Recognition, 2018, p. accepted.  
380
- [28] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv  
preprint arXiv:1412.6980.
- [29] G. Schwarz, et al., Estimating the dimension of a model, The annals of  
statistics 6 (2) (1978) 461–464.
- [30] J. C. Gower, Encyclopedia of Statistical Sciences, Vol. 5, John Wiley &  
385 Sons, New York, 1985, Ch. Measures of similarity, dissimilarity and dis-  
tance, pp. 397–405.
- [31] J. C. Gower, Some distance properties of latent root and vector methods  
used in multivariate analysis, Biometrika 53 (3-4) (1966) 325–338.

- 390 [32] S. Calinon, F. D'halluin, E. L. Sauser, D. G. Cakdwell, A. G. Billard, Learning and reproduction of gestures by imitation, in: International Conference on Intelligent Robots and Systems, 2004, pp. 2769–2774.

## 12.8 Talking with Sentiment: Adaptive Expression Generation Behavior for Social Robots

**Title:** Talking with Sentiment: Adaptive Expression Generation Behavior for Social Robots

**Authors:** I. Rodriguez, Adriano Manfrè, Filippo Vella, I. Infantino, Elena Lazkano

**Conference:** International Workshop of Physical Agents (WAF)

**Publisher:** Springer

**DOI:** 10.1007/978-3-319-70022-9\_66

**Year:** 2018

# Talking with Sentiment: Adaptive Expression Generation Behavior for Social Robots

Igor Rodriguez<sup>1</sup>, Adriano Manfré<sup>2</sup>, Filippo Vella<sup>2</sup>, Ignazio Infantino<sup>2</sup> and Elena Lazkano<sup>1</sup>

**Abstract**—This paper presents a neural-based approach for generating natural gesticulation movements for a humanoid robot enriched with other relevant social signals depending on sentiment processing. In particular, we take into account some simple head postures, voice parameters, and eyes colors as expressiveness enhancing elements. A Generative Adversarial Network (GAN) allows the proposed system to extend the variability of basic gesticulation movements while avoiding repetitive and monotonous behavior. Using sentiment analysis on the text that will be pronounced by the robot, we derive a value for emotion valence and coherently choose suitable parameters for the expressive elements. In this way, the robot has an adaptive expression generation during talking. Experiments validate the proposed approach by analyzing the contribution of all the factors to understand the naturalness perception of the robot behavior.

## I. INTRODUCTION

Social robots represent a great research challenge, aiming to an effective introduction in human everyday life of intelligent embodied machines. Robot social capabilities require both a deep understanding of human behavior and acting with naturalness during the interaction with humans [10]. Naturalness means that a human user could have similar perceptive inputs while interacting with other people considering both verbal and non-verbal signals, social and cultural context, subjectiveness and psychological effect as empathy, emotional impact, and so on.

The gestures, postures, and movements of the body and face expressions are used to convey information about the emotions and thoughts of the sender while supporting verbal communication. Body language represents the key to express feelings and helps the people to understand sociability [13]. McNeill [15] distinguishes four major types of gestures by their relationship to the speech: deictic, iconic, metaphoric, and beats. Unlike the others types, beats are not associated with particular meanings, and they occur with the rhythm of the speech. Such kind of gestures have been considered in this work. While speaking, the robot has to generate credible body language that should shape and convey the information content. It can be derived and learned from humans so that it is consistent with socio-cultural expectation of the interlocutor. Moreover, the robot has to own an emotive

model to dynamically drive the interaction and to establish a relevant emotional link with the interlocutor.

The main contribution of this work is the development of a robot behavior that endows humanoid robots with the ability to generate natural gesticulation movements enriched with several social signals depending on the sentiment of the speech. A Generative Adversarial Network (GAN) allows the proposed system to extend the variability of some basic gesticulation movements avoiding repetitive and monotonous behavior. Furthermore, we take into account some simple head postures, voice parameters and eye LEDs colors to enhance the expressiveness of humanoid robots. Two different experiments have been performed with people in front of a SoftBank's Pepper robot showing our adaptive expression generation behavior. Experiments validate the proposed approach by analyzing the contribution of each of the factors to understand the naturalness perception of the robot behavior.

## II. RELATED WORK

Social robotics [4] aims to provide robots with artificial social intelligence to improve human-machine interaction and to introduce them in complex human contexts. The demand for robot's sophisticate behaviors requires to model and implement human-like capabilities to sense, to process, and to act/interact naturally by taking into account emotions, intentions, motivations, and other related cognitive functions. In recent years a lot of effort has been put in trying to make those behaviors convey sentiment. Several works propose facial expressions as principal mechanism to show emotions, but there are also other possible communicative channels that can be easily understood by a human. For example, colors can be dynamically associated with emotions by suitable cognitive models [2][11]. Low-resolution RGB-LEDs can evoke associations to basic emotions (happiness, anger, sadness, and fear), by using suitable colors and dynamic light patterns [7]. Johnson et al. [12] investigate how LED patterns around the eyes of Softbanks NAO robot can be used to imitate human emotions.

As a matter of fact, postures and movements are relevant for social interactions even if they are subjective and culture dependent. During verbal communication, the level of trust of the human with respect to the robot is higher when the robot's gaze is in the direction of the interlocutor [18]. In [1], authors propose a multimodal robot behavior, expressed through speech and gestures, in which the robot adapts its behavior to the interacting human's personality, and they explore the perception of the interacting human comparing

\*This work was not supported by any organization

<sup>1</sup> I. Rodriguez is with Faculty of Informatics, Computer Sciences and Artificial Intelligence, University of Basque Country (UPV/EHU), Manuel Lardizabal 1, 20018 Donostia, Spain igor.rodriguez@ehu.eus

<sup>1</sup> A. Manfre is with Institute for High Performance Computing and Networking (ICAR), National Research Council of Italy (CNR), Palermo, Italy adriano.manfre@icar.cnr.it

the multimodal behavior with the single-modal behavior, expressed only through speech.

In the field of computer graphics and virtual agents, the results obtained on realistic animation of humans are impressive and allow designers to animate characters with movements by synchronizing nonverbal behavior with synthesized speech [16] [5]. Naturally, speech plays a relevant role to convey emotions, and human voice can be shaped in a very complex way. In the context of human-robot interaction, Crumpton and Bethel explain the importance of using vocal prosody in robots to convey emotions [6].

With respect to the aim of the present work, to the best of our knowledge, the literature does not provide an approach that tries to combine all the previous aspects in a social robot. In the following, we introduce our approach to generate an adaptive expression behavior for social robots.

### III. SYSTEM ARCHITECTURE

In this section, the architecture of the expression generation behavior, which we have named “Adaptive Talking Behavior”, is described. The expression generation process can be summarized in three main steps:

- 1) Extract the sentiment from the text. A sentiment analyzer assesses the sentiment of the text and gives as output a descriptor with information about the polarity of the sentiment (positive/negative/neutral).
- 2) Sentiment to emotion conversion. In this step, the sentiment polarity is encoded into emotion. Only sadness, happiness and neutral emotions have been considered in this work.
- 3) Generate the appropriate expression. The translation from emotion into expression is performed in this stage. The robot shows an emotional expression by means of body expression (talking gestures), facial expression (eyes lighting), and voice intonation (pitch and speed variation).

Those three steps are further detailed in sections IV, V, and VI, respectively.

This “Adaptive Talking Behavior” is composed by several ROS (Robot Operating System)<sup>1</sup> modules, as illustrated in figure 1. The “Sentiment Analyzer” analyzes the sentiment of the text, the “Emotion Selector” converts the sentiment into an emotion the “Eyes Lighting Controller” manages the eyes color the “Speech Synthesizer” tunes voice parameters, and the “Gestures Generator” generates the body expression including arms, hands and head.

### IV. TEXT SENTIMENT EXTRACTION

Sentiment analysis is the research field related to the analysis of people’s opinions, sentiments, evaluations, attitudes, and emotions from written language [17]. The main purpose of sentiment analysis is to extract the polarity (positive/negative/neutral) of a given text.

In order to extract the sentiment from the text we use the VADER sentiment analyzer [9], a lexicon and rule-based

sentiment analysis library that analyzes the polarity and the intensity of sentiments expressed in social media contexts, also generally applicable in other domains. VADER, which is based on dimensional affective models, gives an output composed by: (1) the score ratios for proportions of text that fall in each category, and (2) a compound score, obtained by summing the *Valence* scores of each word in the lexicon (see section V for a deeper explanation about dimensional affective models).

For the phase described above, we have developed a ROS module, named “Sentiment Analyzer”, which takes as input what the robot is going to say (a text) and gives as output the sentiment polarity (negative/neutral/positive) and the compound score obtained by using the VADER sentiment analyzer.

### V. SENTIMENT TO EMOTION CONVERSION

Dimensional affective models represent affective experiences according a set of interrelated and ambiguous states [19]. Emotions are described as linear combinations of *Valence-Arousal-Dominance* (VAD). *Valence* defines how positive or negative the stimulus is, *Arousal* specifies the level of energy and *Dominance* defines how approachable the stimulus is.

The *Valence* deals with the positive or negative character of the emotion, which scales from sadness to happiness. Taking into account that the compound score provided by the VADER sentiment analyzer is obtained from the *Valence* scores of the words and then normalized between  $-1$  to  $+1$  (from most negative to most positive), we compute a conversion from sentiment to emotion through a direct translation of the compound score into the sadness-happiness continuum in the *Valence* axis (Happiness: compound score  $\geq 0.5$ , Neutral: compound score  $> -0.5$  and compound score  $< 0.5$ , Sadness: compound score  $\leq -0.5$ ).

The conversion from sentiment to emotion is done by the “Emotion Selector” ROS module, which takes as input the result obtained from the “Sentiment Analyzer”. For the time being, the emotion appraisal is done as a direct translation from the sentiment value into emotion in sadness-happiness continuum. It is worth mentioning that more inputs and more emotions should be considered for the emotion appraisal in the future.

### VI. EXPRESSION GENERATION

Emotion expression is one of the characteristics that make us social beings. It allows us to communicate our emotional state and, at the same time, it gives us a glimpse into the inner mental state of other individuals. Emotional expressions can occur with or without self-awareness during both verbal and nonverbal communication, and can be manifested in different ways, such as facial movements, body postures, gestures, etc.

Our approach to appropriately express the emotion obtained from the sentiment processing of the text consists of: mapping the emotion into expression that combines natural body gestures, enriched with facial expressions and voice

<sup>1</sup><http://wiki.ros.org/>

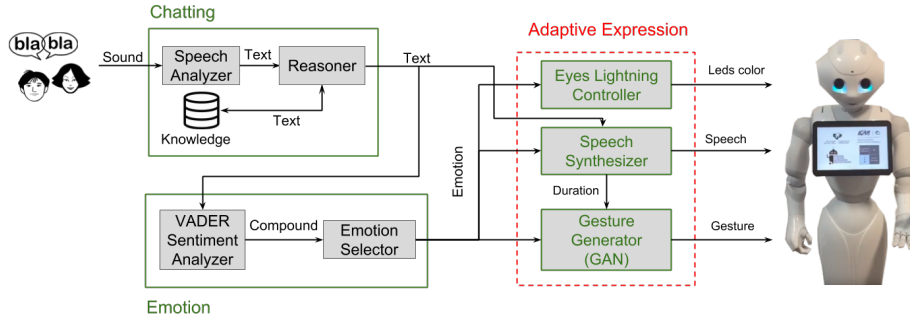


Fig. 1. Description of the Adaptive Talking Behavior architecture

intonation. Several video examples about Pepper talking with sentiment are available in RSAIT’s YouTube channel<sup>2</sup>.

### A. Gestures Generation

Typically, humans use gestures (head, hands and arms movements) to communicate with others; gestures are used both to reinforce the meaning of the words and to express feelings through non-verbal signs. How an emotion is reflected in the different parts of the body is well explained in [14]. Humanoid robotics platforms, like Pepper, help us to explore body language expression approaches. Thanks to their high expression capabilities we can build natural robot behaviors that enhance the expressiveness of robots. In a previous work [20], Rodriguez et al. propose a robot behavior that makes a NAO robot able to talk and gesticulate showing some emotions. The robot executes randomly selected predefined gestures adapting the execution tempo of the movements according to the sentiment of the text.

The short set of gestures (default animations in the NAOqi API<sup>3</sup>) that Pepper and NAO have for gesticulating while speaking make their expression ability limited and repetitive. In order to overcome this limitation, our approach based on Generative Adversarial Networks (GAN) enables humanoid robots to dynamically generate synthetic gestures (composed by arm, hand and head movements) during verbal communication at run-time.

Generative Adversarial Networks [8] are semi-supervised emerging models that basically learn how to generate synthetic data from the given training data. A GAN network is composed by two different interconnected networks. The Generator ( $G$ ) network generates possible candidates so that they are as similar as possible to the training set. The second network, known as Discriminator ( $D$ ), judges the output of the first network to discriminate whether its input data are “real”, namely equal to the input data set, or if they are “fake”, that is generated to trick with false data. The general architecture of a GAN with  $G$  and the  $D$  networks is shown in Figure 2.

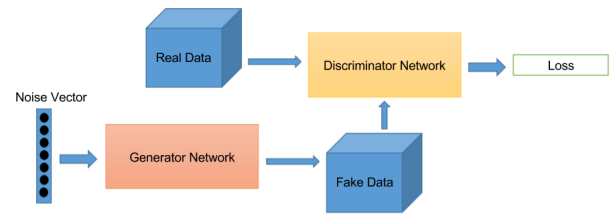


Fig. 2. Description of GAN architecture

In the first step, the  $D$  takes as input both, real data and fake data, and returns for each sample its probability to be real or not. In the second step, the  $G$  network is trained. While the parameters of  $D$  are fixed, in each epoch, the weights of the  $G$  network are updated to let the discrimination results on the sample generated by  $G$  as near as possible to 1. That is, this second step is aimed to modify the  $G$  network in order to be able to generate samples that can trick the  $D$  network.

The  $G$  network is never exposed to the real data, the only manner to enhance its generation capability is through the interaction with  $D$  by means of the output. Instead,  $D$  has access to both, real data and fake data, and produces as output the ground truth to know if the data came from the generator or the dataset. The discriminator’s output value is exploited by the generator to enhance the quality of the forgeries data.

Back-propagation is applied in both networks to enhance the accuracy of the generator to produce valid movements; on the other side, the discriminator becomes more skilled to false flag data. In this work we use a batch size of 16 and we trained the GAN network for 2000 epochs.

The training dataset given to the  $D$  network to learn the distribution space of the data is composed by the gestures obtained from the default animations of the NAOqi API. We have chosen a subset of those gestures that can be used for accompanying the speech, and can be performed individually or in composition to constitute complex sequences of gestures. On the other hand, the Generator is seeded through a random input with a uniform distribution in the range  $[-1, 1]$ . The Generator produces as output gestures that

<sup>2</sup><https://www.youtube.com/channel/UCT1s6oS21d8fxFeugxCrjnQ>

<sup>3</sup><http://doc.aldebaran.com/2-5/ref/python-api.html>

belong to the real data distribution and that the D network is not able to evaluate as generated or real.

In order to extend the variability of the gestures, we sampled the default NAOqi animations with a frequency of 4 Hz, obtaining a set of 1500 robot poses. A pose is represented by a set of 14 joint values, comprising robot’s head, hands and arms. The composition of four consecutive poses is a gesture (movement segment) that is one instance of the training set. The GAN network could have been trained using poses instead of movement segments, but then, the outputs would be single poses that should be afterwards concatenated to generate talking movements. However, this approach would produce less smooth gestures.

The talking gestures generation is done by the “Gestures Generator” ROS module. that takes as input the emotion value obtained from the “Emotion Selector” and produces a number of gestures that are well suited to be executed with the right velocity according to the speech duration. The execution velocity is influenced by the recognized emotion in order to express better the feeling, i.e. if the emotion to be shown is “happy” the gesture will be executed at a faster pace than whether to gestures to be performed is bound to the emotion “sad”.

Also the head tilt is influenced by the emotion. If the emotion is neutral, the robot will look forward. However, if the emotion is happy the robot will tilt the head upwards, or downwards if it is sad. The emotion *Valence* obtained by the emotion appraisal is normalized between the maximum and minimum values for the head tilt. Figure 3 shows an example of gestures generated using GAN for sadness, neutral and happiness emotions.

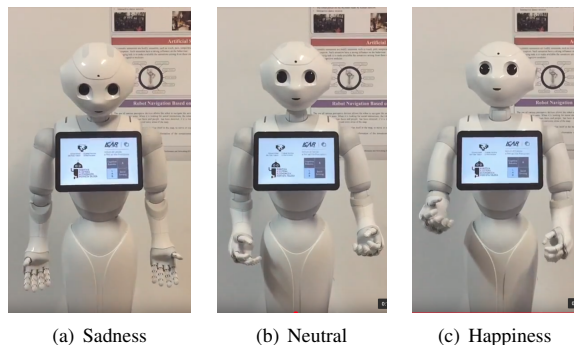


Fig. 3. Some examples of generated gestures with emotion

### B. Facial Expression

The design of humanoid robots’ eyes is usually inspired by human face, trying to exactly reproduce human eyes’ shape and movements. However, SoftBank’s robots have some limitations due to the structure of their eyes. In particular, Pepper robots’ eyes are composed by two rings of LEDs with a black pupil inside. The LEDs can be controlled to show different hues, change color intensity and can be turned on/off for different time duration.

Johnson et al. [12] demonstrate in their work that NAO’s eyes can be used to express emotions. Taking inspiration from their color-emotion study, in our approach we adopt the same color configuration, and in addition we use the emotion *Valence* to change the intensity of the color.

The “Eyes Lighting Controller” is employed to convert emotion into facial expression, specifying the color and the intensity of each eye LEDs (see figure 4). The controller exploits the emotion *Valence* value to codify it into RGB space to be displayed in the robot.

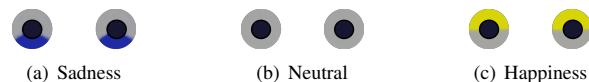


Fig. 4. Sadness: blue-greenish color from RGB(0, 0, 255) to RGB(0, 255, 255). Neutral: no color from last RGB to RGB(0, 0, 0). Happiness: yellow color from RGB(76, 76, 0) to RGB(255, 255, 0).

### C. Voice Intonation

In ordinary life humans use different voice intonation depending on the context in which they are and also to emphasize the message being conveyed. The voice intonation has a key role to understand the mood of the speaker. The influence of the voice intonation in emotional expression is clearly argued in [3]. The authors prove that some emotions, such as fear, happiness and anger, are portrayed in a higher speech rate and also at a higher pitch than emotions such as sadness.

We have used the happiness, neutral and sadness intonations to portray the three emotions available in our system. Unfortunately, Pepper’s speech synthesizer does not offer direct voice intonation selection, but it provides the option to setup voice parameters such as pitch and speech rate, which can be tuned to obtain a different voice intonation than the standard provided. Our approach consists of changing the pitch and speed rate parameter values according to the emotion *Valence* value, i.e. the emotion’s *Valence* obtained from the emotion appraisal is normalized between the maximum and minimum values for the voice pitch and speed rate. Maximum and minimum values have been experimentally defined for our system.

## VII. EXPERIMENTAL SETUP

In this section we introduce the robotic platform used in the experiments, the hypotheses we want to validate through the experiments, and an overview for the conducted experiments.

### A. Robotic Platform

The robotic platform employed in the performed experiments is a Pepper robot developed by Softbank Robotics<sup>4</sup>. Pepper has a height of 120cm and 20 degrees of freedom human-like torso that is fitted onto a wheeled platform, equipped with full-color RGB LEDs (placed in eyes, ears

<sup>4</sup><https://www.ald.softbankrobotics.com/en/robots/pepper>

and shoulders), three cameras, and several sensors located in different parts of its body that allow for perceiving the surrounding environment with high precision and stability.

### B. Hypothesis

The presented research aims to test and validate the following hypotheses:

- H1: Generative Adversarial Networks can be used to generate gesticulation movements and the gestures generated will be considered as natural by the user.
- H2: The robot expression displayed through generated body gestures adapted and combined with eyes' colors and voice intonation are perceived as more expressive by the user than expressing only through talking arm gestures.

### C. Experimental design

In order to test and validate those two hypothesis (H1 and H2), we have defined two different experiments (E1 and E2) in which participants must judge the robot behavior by filling a questionnaire designed to analyze their perception about the system.

- E1: The robot reports short news to the participants. It repeats three times the same piece of news using a different type of gesticulation in each session: using *Random* mode of Softbank's Animated Speech module, which launches some neutral animations executed one after another; using *Contextual* mode of Softbank's Animated Speech module, which launches some specific animations each time a keyword is detected, and when no contextual animations is found, it randomly launches a new animation; using gestures generated by GAN.
- E2: The robot tells different tweets to the participants. It repeats three times the tweets, but adding a new feature to the robot expression each session. First, talking gestures are generated adding head tilt and arms movement speed features (S1). In the next session (S2), the color of the eyes is added together with the features added in S1. Finally, in the last session (S3), the tone of the voice is added together with the features added in S1 and S2.

## VIII. EXPERIMENTAL RESULTS

For the evaluation of the system, 57 voluntaries have been recruited at ICAR-CNR in Palermo, Italy and the University of the Basque Country (UPV/EHU) in Donostia, Spain, to judge the behavior of Pepper during a talk. The participants grouped into three or four, entered in the experiment room without any information about the experiments and were seated in front of the robot.

In the first experiment (E1), participants evaluated the talking gestures performed by the robot. The order of the type of gestures performed by the robot was randomized in order to avoid possible bias of people always choosing the last remembered as best behavior. After seeing each session participants filled a questionnaire based on five-point Likert scale rating the following aspects: the *naturalness* (A)

of the gestures, the *fluency* (B), the *appropriateness* of the gestures for accompanying the speech (C), the *variability* of gestures perceived (D), the *synchronization* (E) between the speech and the gestures, and how much they *liked* (F) the gestures performed by the robot. Additionally, after seeing all sessions, they chose their preferred one.

	Random	Contextual	GAN
(A) Naturalness	3.4	3.4	3.2
(B) Fluency	3.5	3.6	3.3
(C) Appropriateness	3.2	3.2	3.2
(D) Variability	3.1	3.1	3.0
(E) Synchronization	3.3	3.3	3.3
(F) Liking	3.4	3.4	3.3

TABLE I  
MEAN VALUES FOR EACH GESTICULATION TYPE.

Results in table I show that the GAN approach allows the system to produce credible movements during the speech even if no contextual information is used. The mean scores for each gesticulation type are very similar. Nevertheless, when asking about preferred session GAN based approach obtained 41%, while Random and Contextual obtained 30% and 29% respectively. The main advantage of the GAN approach is that the robot can use different datasets of simple movements depending on cultural context, social practices, or individual preferences.

In the second experiment (E2) the potentiality of the generated robotic gesticulation (by using GAN) is evaluated considering the possibility to convey also emotions depending on the pronounced speech, and enriching the interaction with other relevant factors such as the head movements and arms movement speed, the tone of the voice, and the color of the eyes. Participants must judge the gestures in the same way as in the first experiment (questions A-F), but they also must identify in which *part of the body* they appreciated the emotion (G), and how much they *liked* the overall robot capability to perform expressions (H).

Results in figures 5, 6, and 7 show that robot's expressiveness improves significantly by adding more features as eyes LEDs colors and voice tone. In particular, results of table II show the influence of each relevant part in the different sessions: participants clearly appreciate the effect of head tilt when expressing emotions, and they also perceived the influence of changing the arms movement speed according to the emotion; the introduction of the eye LEDs colors during S2 has a great impact in expressiveness perception; also the modulation of the voice tone is well perceived in S3.

(G) Relevant part	S1	S2	S3
Head	79%	64%	71%
Arms	43%	36%	46%
Eyes	2%	80%	70%
Voice	16%	14%	57%

TABLE II  
INFLUENCE OF EACH RELEVANT PART IN THE DIFFERENT SESSIONS.



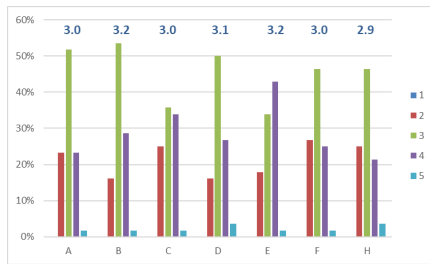


Fig. 5. Results of E2-S1: GAN + Head tilt + Arm velocity. Mean scores are reported above each histogram.

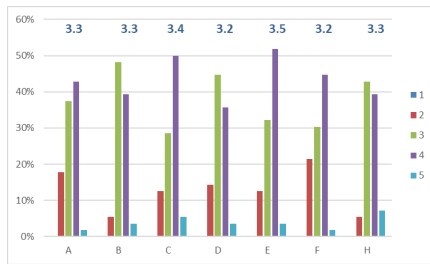


Fig. 6. Results of E2-S2: S1 features + Eye LEDs colors

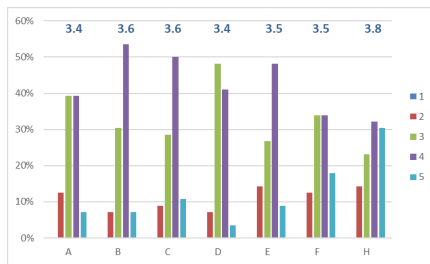


Fig. 7. Results of E2-S2: S1 and S2 features + Voice tone.

## IX. CONCLUSIONS AND FURTHER WORK

A suitable generator of rhythmic gesticulation movements with few others expressive features (the head posture, the arm movements velocity, the variation of voice tone, the change of color of eyes) dependent from sentiment detected on sentences, could be a simple system to have high appreciation rates during a human-robot conversation. On the basis of the obtained results, the further work could focus the following interesting directions. The dataset of basic movements used by the GAN approach could be derived from the direct observation of the human by using an RGBD device. In this way, a robot can use with a given person a set of movements that are familiar to him/her. Furthermore, we plan to introduce the detection of the six basic emotions to have a more complex expressive behavior and to consider also some gestures with metaphoric or iconic meaning. We will perform similar experiments exploiting the two different cultural contexts (Italian and Spanish) aiming to investigate for instance the effects of the cultural influences. Moreover, we should investigate to find also evaluation methods to measure the goodness of the gestures generated by the GAN

network.

## ACKNOWLEDGMENT

This work has been partially supported by the Basque Government (IT900-16), the Spanish Ministry of Economy and Competitiveness MINECO/FEDER (TIN 2015-64395-R, MINECO/FEDER,EU). Author Rodriguez has received the UPV/EHU research Grant PIF13/104.

## REFERENCES

- [1] Amir Aly and Adriana Tapus. Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human-robot interaction. *Autonomous Robots*, 40(2):193–209, 2016.
- [2] Agnese Augello, Ignazio Infantino, Giovanni Pilato, Riccardo Rizzo, and Filippo Vella. Binding representational spaces of colors and emotions for creativity. *Biologically Inspired Cognitive Architectures*, 5:64 – 71, 2013.
- [3] Tanja Bänziger and Klaus R Scherer. The role of intonation in emotional expressions. *Speech communication*, 46(3):252–267, 2005.
- [4] Cynthia Breazeal. *Designing sociable robots*. Intelligent Robotics and Autonomous Agents. MIT Press, Cambridge MA, USA, 2004.
- [5] Justine Cassell, Hannes Högni Vilhjálmsón, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486. ACM, 2001.
- [6] Joe Crumpton and Cindy L. Bethel. A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics*, 8(2):271–285, Apr 2016.
- [7] Johannes Feldmaier, Tamara Marmat, Johannes Kuhn, and Klaus Diepold. Evaluation of a RGB-LED-based emotion display for affective agents. *arXiv preprint arXiv:1612.07303*, 2016.
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [9] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.
- [10] Ignazio Infantino. Affective human-humanoid interaction through cognitive architecture. In *The Future of Humanoid Robots - Research and Applications*. Dr. Riadh Zaier (Ed.), InTech, 2012.
- [11] Ignazio Infantino, Giovanni Pilato, Riccardo Rizzo, and Filippo Vella. I feel blue: Robots and humans sharing color representation for emotional cognitive interaction. In *Biologically Inspired Cognitive Architectures 2012*, pages 161–166. Springer, 2013.
- [12] David O Johnson, Raymond H Cuijpers, and David van der Pol. Imitating human emotions with artificial facial expressions. *International Journal of Social Robotics*, 5(4):503–513, 2013.
- [13] Heather Knight. Eight lessons learned about non-verbal interactions through robot theater. *Social Robotics*, pages 42–51, 2011.
- [14] Margot Lhomme and Stacy Marsella. *The Oxford Handbook of Affective Computing*, chapter Expressing Emotion Through Posture and Gesture, pages 273–285. Oxford University Press, 2015.
- [15] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.
- [16] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. Graph.*, 27(1):5:1–5:24, March 2008.
- [17] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.
- [18] Raul Benites Paradedá, Mojgan Hashemian, Rafael Afonso Rodrigues, and Ana Paiva. How facial expressions and small talk may influence trust in a robot. In *International Conference on Social Robotics*, pages 169–178. Springer, 2016.
- [19] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.
- [20] Igor Rodriguez, José María Martínez-Otzeta, Elena Lazkano, and Txelo Ruiz. Adaptive emotional chatting behavior to increase the sociability of robots. In *International Conference on Social Robotics*, pages 666–675. Springer, 2017.

## 12.9 Robots on Stage: a Cognitive Framework for Socially Interacting Robots

**Title:** Robots on Stage: a Cognitive Framework for Socially Interacting Robots

**Authors:** I. Rodriguez, A. Astigarraga, E. Lazkano, J. M. Martínez-Otzeta, I. Mendiàdua

**Journal:** Biologically Inspired Cognitive Architectures (BICA)

**Publisher:** Elsevier

**DOI:** 10.1016/j.bica.2018.07.014

**Year:** 2018



Contents lists available at ScienceDirect

## Biologically Inspired Cognitive Architectures

journal homepage: [www.elsevier.com/locate/bica](http://www.elsevier.com/locate/bica)

## Research article

## Robots on stage: A cognitive framework for socially interacting robots

Igor Rodriguez\*, Aitzol Astigarraga, Elena Lazkano, José María Martínez-Otzeta, Inigo Mendialdua

Department of Computer Sciences and Artificial Intelligence, University of the Basque Country UPV/EHU, Donostia-San Sebastian 20018, Spain



## ARTICLE INFO

## Keywords:

Social robotics  
Affective perception  
Cognitive architecture

## ABSTRACT

This article is an attempt to characterize the cognitive skills involved in the development of socially interacting robots. We argue that performative arts, such as oral improvised poetry, can serve as a useful testbed for the development and evaluation of robots that interact with humans. The paper presents a speech-based humanoid poet-performer that can (1) listen to human commands and generate poems on demand; (2) perceive audience's feedback and react displaying the corresponding emotional response; and (3) generate natural gesticulation movements enriched with social signals depending on sentiment processing. We discuss each of the involved abilities, present working implementations and show how they are combined in an embodied cognitive architecture to achieve the fluent coordination and joint-action timing needed in live events.

## Introduction

Social robotics (Breazeal, 2004) aims to provide robots with artificial social intelligence to improve human–machine interaction and to introduce them in complex human contexts. The demand for robot's sophisticated behaviors requires to model and implement human-like capabilities to sense, to process, and to act/interact naturally by taking into account emotions, intentions, motivations, and other related cognitive functions. And, of course, the ability to communicate through natural language and non-verbal signs is in the front line of research.

Nowadays, the development of control architectures for robots while taking into account the complexity of social human-robot interaction is a real challenge. It requires various cognitive features to be present: emotions, attention allocation, creativity and reactive and deliberative levels of perception and action.

Ever since the pioneering research on cognitive architectures (Newell, 1994), several architectures can be found in literature: SOAR (Laird, Kinkade, Mohan, & Xu, 2012), ACT-R (Anderson, 2005), CLARION (Sun, 2006), iCub (Vernon, Metta, & Sandini, 2007a) and ICARUS (Choi & Langley, 2018), among others. A good review of the literature can be found in (Vernon, Metta, & Sandini, 2007b; Thórisson & Helgasson, 2012; Langley, Laird, & Rogers, 2009). But, in spite of the numerous contributions in the field of cognitive architectures, robots that can listen to human speech, understand it, interact according to the conveyed meaning and respond still represent major research and technological challenges. Therefore, a research on different approaches to build control architectures oriented for interaction able to deal with

cognitive capabilities such as emotion and social aspects of human-robot interaction (HRI) is highly useful.

In the last years research in the field of social robotics with conversational capabilities has grown up, and several robots have been designed and developed in this area. Such applications, most of them not concerned about being a faithful model of cognition, comprise several cognitive abilities and provide robust adaptive behaviour for human-robot interaction. Relevant works include: industrial and manufacturing robots (Cherubini, Passama, Crosnier, Lasnier, & Fraise, 2016; Heyer, 2010); assistive robots and robots focused on aiding users with special needs (Bemelmans, Gelderblom, Jonker, & De Witte, 2012; Fasola & Mataric, 2012; Gómez Esteban et al., 2016; Kachouie, Sedighadeli, Khosla, & Chu, 2014; Luria, Hoffman, Megidish, Zuckerman, & Park, 2016; Tapus, Tapus, & Mataric, 2009); interactive teachers and educational assistants (Fridin & Belokopytov, 2014; Kanda, Shimada, & Koizumi, 2012); lab or household robotic assistants (Dautenhahn et al., 2005; Wisspeintner, Van Der Zant, Iocchi, & Schiffer, 2009); shopping mall guides (Chen et al., 2015); persuasive robots (Chidambaram, Chiang, & Mutlu, 2012; Lee & Liang, 2016); museum robots (Kanda, Arai, Suzuki, Kobayashi, & Kuno, 2014; Rashed, Suzuki, Lam, Kobayashi, & Kuno, 2015) and tour guides (Kanda et al., 2014); companion robots (Moyle et al., 2013); and robots more oriented to the entertainment area, such as robotic theater actors (Fernandez & Bonarini, 2013; Hoffman, 2011), musicians, storyteller robots (Bruce, Knight, Listopad, Magerko, & Nourbakhsh, 2000; Bae et al., 2012; Costa, Brunete, Bae, & Mavridis, 2016; Wu, Wang, Tay, & Wong, 2017) and dancers (Kosuge, Hayashi, Hirata, & Tobiyama,

\* Corresponding author.

E-mail address: [igor.rodriguez@ehu.eus](mailto:igor.rodriguez@ehu.eus) (I. Rodriguez).<https://doi.org/10.1016/j.bica.2018.07.014>Received 12 June 2018; Received in revised form 11 July 2018; Accepted 11 July 2018  
2212-683X/ © 2018 Elsevier B.V. All rights reserved.

2003). We could denominate this last group as performance robots, that is, robots close to performative arts that execute their task on a stage. An approach that brings the gap between cognitive models and social robotics is presented in Augello, Infantino, Pilato, Rizzo, and Vella (2015) and Augello et al. (2016). In those works authors propose a cognitive architecture for computational creativity, making a humanoid robot able to dance.

We believe that stage performance is valuable both as an implementation platform and as a testing ground for interaction-oriented cognitive architecture research. On the one hand, the event setting is constrained to some degree, limiting thus the perception and actuation possibilities of the robotic system. On the other hand, it provides a unique environment in which humans and robots collaborate incorporating dialog, sensory processing, action selection and behavior coordination.

Our robotic system, called Bertsobot, should be framed within performance robots: an autonomous robot that participates on live events, improvising poems under given constraints and performing them on stage. Thus, Bertsobot brings together capabilities and characteristics from many of the previously mentioned performance robots: theatrical staging, verbal and non-verbal communication, people detection and key stage elements perception, affect detection and emotional response, timing and coordination, etc.

In this article we present working implementations of the involved cognitive skills and show how they are combined to achieve the fluent coordination and joint-action timing needed in live events.

We do not claim to address the issue as a whole. This article attempts however to organize it into a coherent challenge for social robotics, and to explain and illustrate some of the paths that we have investigated on our robots, which result in a robot architecture designed for human-robot interaction that implements cognitive skills.

### Improvised poetry and *Bertsolaritza*

Writing poetry requires both creativity to construct a meaningful message and lyrical skills to produce rhyme patterns and follow metrical constraints. Furthermore, oral poetry, poetry constructed without the aid of writing (Lord, Mitchell, & Nagy, 2000), implies that a work has to be composed and performed at the moment, with no prior preparation. Nowadays many improvisational oral practices exist around the world, such as Serbo-Croatian *guslars* (Lord et al., 2000), freestyle rap (Pihel, 1996) and Basque *bertsolaritza* (Garzia, Sarasua, & Egaña, 2001).

*Bertsolaritza*, the art of improvising verses in *Euskara* (the language of the inhabitants of the Basque Country) is one of the manifestations of traditional Basque culture that is still very much alive. Events and competitions in which the verse-makers, *bertsolari-s*, have to produce impromptu compositions about topics or prompts are very common. A typical scenario involves an emcee suggesting a topic to the *bertsolari*, who must then, within the space of less than a minute, come up with a verse on that topic that must obey certain rules; in other words, it must fit in with a prescribed verse-form that also involves a rhyme scheme and a melody (chosen from among hundreds of tunes). And of course perform that verse, before an audience and without any musical accompaniment (see Fig. 1).

### The Bertsobot cognitive architecture

The Bertsobot system endows the robots with some of the *bertsolari-s*' capabilities that allow to take part in public performances. Therefore, our Bertsobot system is able to perceive the feedback and emotions of the audience through their applause and react accordingly, as human oral improvisers do, modifying in real time the sentiment of the poem and its corporal expression accordingly. We focused on creating a practical cognitive architecture that follows the dynamics of real events, as verse-makers do:



Fig. 1. Typical scenario.

1. Wait sitting for its turn.
2. When demanded, place itself in front of the microphone and listen to the exercise proposed by the emcee.
3. Compose and sing the verse to the public.
4. Observe and receive audience's feedback and react accordingly.
5. Go back to its sitting place.

A robot capable of performing the aforementioned tasks involves the development of several cognitive capabilities. Although Bertsobot's main task is to compose verses, which requires a high cognition level, there are other important capabilities that at lower-level manage perceptions and representations of the environment. Specifically, it requires certain abilities to understand verbal instructions, move around the stage, recognize the different key elements of the scenario, interact with other agents and the audience, and show the same degree of expressiveness that *bertsolari-s* show on stage.

The cognitive architecture is the framework that facilitates us the development of cognitive functions, providing a structure within which to embed the mechanisms for perception and action, motivation and social interaction (Vernon, von Hofsten, & Fadiga, 2016). The cognitive description of our robotic system has been inspired by Augello et al. (2018). The framework is suitable to model aspects such as motivation and emotions that are integrated with perceptual and reasoning processes. Fig. 2 shows an overview of the proposed framework for Bertsobot.

In our framework, The Long Term Memory (LTM) stores all the knowledge required by the robot to accomplish the task. It contains the postures model with which the robot will be aware of its body configuration, the rules and corpora to generate extemporaneous poems, a set of melodies for singing composed verses, a gesture repertoire related to the expression capabilities of the robot, and a linguistic dictionary. On the other hand, the Short Term Memory (STM) or working memory stores temporal information about robot's and audience's emotional state. Drives comprises all the basic behaviors to interact with the environment and extract information from it. Finally, the Social Interaction module guides the human-robot interaction through verbal communication and body expression.

Our cognitive framework has been designed as the basis for the development of an adaptive robot for human-robot interaction, integrating a wide range of components in a scalable ROS<sup>1</sup> based control architecture. It is composed by different behaviours or modules that make the robot act in a consistent manner and resemble a real *bertsolari*.

In the following sections, we introduce in detail the main cognitive capabilities used in our robot performer.

<sup>1</sup> <http://www.ros.org/>.

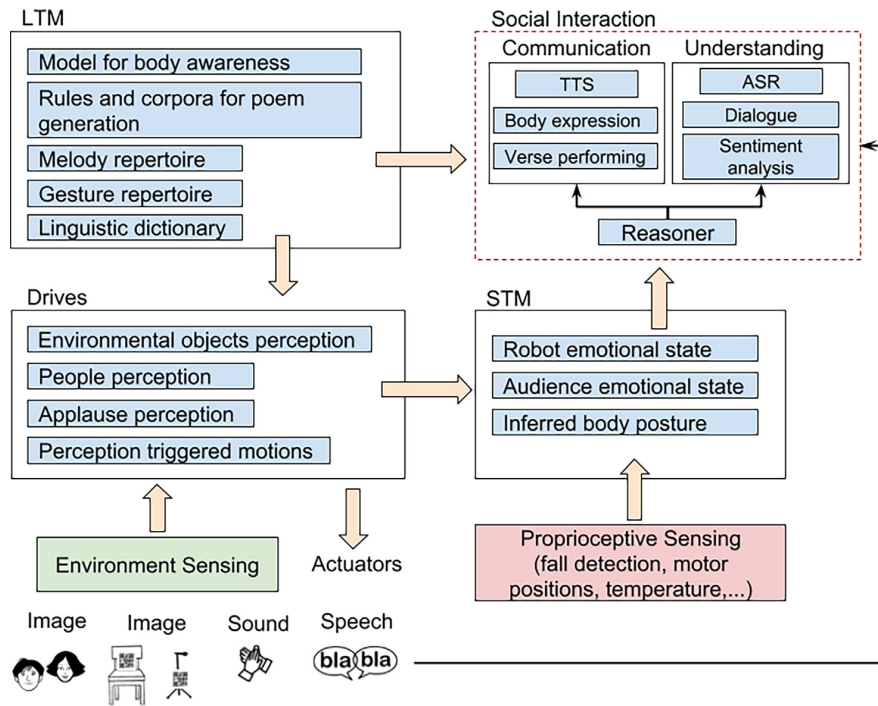


Fig. 2. Bertsobot's cognitive architecture.

**Verbal communication**

Verbal communication is essential if we intend fluid HRI. Speech-based interaction is accomplished in two ways in our Bertsobot system: on the one hand, the system is able to maintain a dialogue with its interlocutor to receive instructions about the stage performance: when to start and when to finish, the theme and the metric to compose the poem and so on. Moreover, speech acts should be coordinated with the robot physical movement in space (whether that movement is functional or expressive). On the other hand, oral communication is accomplished when the robot creates, under the given instructions, a new piece of poem and sings it with a proper melody.

Fig. 3 shows the architecture of our dialogue system that incorporates an Automatic Speech Recognizer (ASR), Language Interpreter, Dialogue Manager, Response Selector, Text Generator, Speech Synthesizer (TTS) and the Singing Synthesizer (TTSKantari).

The Automated Speech Recognition (ASR) component converts the raw audio input into a sequence of words. Google Speech service is used as an ASR system. This is forwarded to a Language Interpreter module to extract the semantics of the utterance. The Language Interpreter module parses the input text and makes use of a database of keywords to identify user's query. Then, the Dialog Manager (DM) decides upon the action to take according to the employed dialog strategy. In our system, the DM is implemented as a finite state automaton. Thus, the system guides the conversation with the user, asking a series of

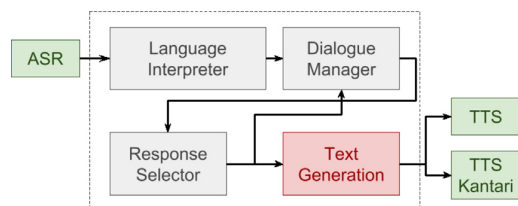


Fig. 3. Description of the Bertsobot dialogue system.

questions and chatting or singing a poem depending on what the user commands. The Response Selector selects the proper text output for the actual state. The output can be a predefined response according to the dialogue's state (a set of utterances to receive information about the stage performance), or a command to compose a novel poem under the given constraints. The Text Generator module receives the input and generates the poem when it is commanded. Finally, the last step converts the text into audio. When the text output is an utterance it is passed to the text-to-speech engine (TTS) component to be synthesized. AhoTTS tool, a speech synthesizer for Basque Language developed by AhoLab (Hernández, Navas, Murugarren, & Etxebarria, 2001) is used for that purpose. And, when the output is a poem text, the last step consists of translating from text to a song that will be immediately performed by the robot. To do so, poem's metric is analysed and a melody according to the sentiment of the text is chosen from an available database. The poem and the melody are sent to the TTSkantari singing synthesizer (Agirrezabal, Alegria, Arrieta, & Hulden, 2012) which produces the audio file with the sung poem.

*Automatic poetry generation*

The core element of the Text Generation module is the Automatic Poetry Generation (Gervás, 2013; Lamb, Brown, & Clarke, 2016; Oliveira, 2017) system. Its goal is to use improvised poems to convey a message and transmit emotions to the public. Our approach implements the same strategy used by *bertsolari-s* for the creation of impromptu verses, and in a few seconds – less than a minute – assembles a new poem along the prescribed verse-form. Although our work focuses on *bertsolaritza*, it can be generalized as automatic poem generation.

The proposed system receives as input the topic of the poem and the affective state (positive, neutral or negative) and tries to give as output a novel poem that: (1) satisfies formal constraints of rhyme and metric, (2) shows coherent content related to the given topic, and (3) expresses them through the predetermined mood. Thus, the system can be asked to view a topic (eg. spring) from a particular affective stance (eg. negative). In doing so, the goal is to not only to convey a message in the



form of a poem but also to respond to an affective target and/or to create an affective response in the audience. That is, creating a poem in an intentional way.

Our poetry generation strategy is a corpus-based method (Astigarraga, Martínez-Otzeta, Rodríguez, Sierra, & Lazkano, 2017) and the overall semantic relationship has been implemented with an LSA model (Astigarraga, Jauregi, Lazkano, & Agirrezabal, 2014; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). The verse generation procedure relies in the extraction of sentences from corpora and combining them (under rhyme and metric constraints) to form the final poem. The LSA model assures the internal coherence between poem lines and the overall coherence with respect to a theme.

The two text generation methods are:

- Sentence retrieval: The basics of this method is to extract from the corpus sentences which meet rhyme and metric constraints.
- N-gram probabilistic model: Starting from the rhyming word, the verse is built backwards using the selected N-gram model; extending at each step the sequence of words with new ones that have a non-zero probability of appearing after the last word.

The overall architecture, depicted in Fig. 4, is modular and provides a high level of customization, depending on the needs of the user.

It must be noted that the sentiment input is intended to receive audience’s feedback. Next section (see Section “Affective perception”) explains how audience reactions are processed and coded. This feedback is used to respond accordingly, maintaining the affective response when the audience reaction is positive, and changing the sentiment target when it is not.

In Table 1 we show two poems generated by the poetry generator system. We also provide an approximate English translation, even though part of its aesthetic value is lost in translation.

Objective evaluation of poetry is difficult, if not impossible, to assess in an automatic way. As Gervás (2015) and Cardoso, Veale, and Wiggins (2009) stated, human evaluators are needed to assess the degree of creativity of a computational creation. Thereby, we contacted with 5 people close to *bertsolaritza* and they participated in the evaluation, explicitly telling them that such poems were the product of an automated system. Each of them analyzed twenty poems, ten from each text generation method. They have been asked to give their overall impression about the overall quality, emotional affect, similarity with the theme, internal coherence and style. General conclusions they extracted:

- The emotional affect of the poems can be clearly appreciated.
- The generated poems are related to the subject. This relationship is not only appreciated through the repetition of the key word or theme, but also through the inclusion in the poem of words semantically similar to the theme.
- Sentence-retrieval method ensures the internal coherence of the sentences (since it extracts entire phrases from the corpus) but, on

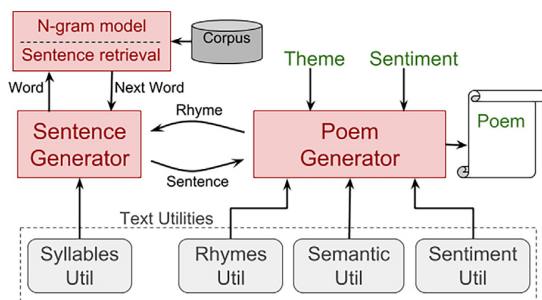


Fig. 4. Poetry generation architecture.

Table 1

Two poems created by the system given theme “music”. The upper one using sentence-retrieval method and with negative sentiment, and the bottom one using the n-gram method and with positive sentiment.

Basque	Gutxien ezagutzen zen musikaria hak bertsoan aurkeztu zuen jaialdia haiengandik aldendu hori nire nahia Bolibarko Txikito puntista ohia
English	The least known of the musicians he introduced the festival improvising verses get away from them, that’s my goal Txikito de Bolivar, ex pelotari
Basque	Zaintzen nahiko lan daukat emanaldiekin eta libre izan nahi dut entzuleekin zorian gehiago gaur ez gaude ezberdin nere musika ez al duzu zuk atsegin
English	I’ve enough taking care of myself in concerts and I want to be free with the listeners more happiness, today we are not different Don’t you like my music?

the other hand, creates more rigid poems.

- The N-gram method is more flexible, malleable, and seems to get closer to the given topic. But, the toll to be paid is that flexibility is sometimes translated into unintelligible phrases.

The final objective of the evaluation was none other than comparing the proposed text-generation methods and measuring the consistency (both internal and external with respect to the given topic). Once the method is implemented in the real robot, the improvised verses will be judged by the public using audience’s applause as feedback, as in the events of human improvisers.

### Affective perception

Audience plays an important role in any type of performances, specially in *bertsolaritza*. The crowd shows how pleasant the verses have been, usually clapping as well as laughing when they have found it amusing. Perceiving and showing emotions is essential to convey interaction. Developing an approach to react to the audience’s feedback covers multiple fields, such as applause detection and classification, and selection of the robots appropriate reaction in the context of the performance. The presented approach uses audience applause as feedback to the robot system (Kraemer, Rodríguez, Parra, Ruiz, & Lazkano, 2016). Applauses are captured and translated into a response from the public by means of energy (E) and duration (d) of the applause. The addressed strategy can be split up into a straight-forward work-flow (see Fig. 5). In the initial step, audio processing and machine learning techniques prepare the input audio stream by first chunking it, and then classifying each chunk as being applause or not. Next, the incoming stream of classified chunks is segmented into sections of consecutive applauses, leading to a small descriptor ([E, d]) for every evaluated applause. Based on all previous applauses of the event, the most recent one can subsequently be classified. The applauses are coarsely categorized as belonging to one of the following classes: Negative, Neutral, Positive and Very Positive.

Therefore, the system detects audience’s feelings and can choose the proper emotional response to display at a cognitive level. Each applause feedback class has been represented with an emotion: Sad, Calm, Joy and Excited emotions correspond to Negative, Neutral, Positive and



Fig. 5. Approach work-flow for applause classification.

Very Positive applause classes (see Section “Gestures” for emotional portrayal details). Right now the robots only maintain a basic persistent emotional state, which endows them to be aware of their current emotions when receiving audience’s feedback, and thus being able to adapt their verse style accordingly.

### Interaction with the environment

Embodied cognition establishes that cognition depends upon experiences that come from having a body and thus, feedback between agents and the world is essential to develop cognitive capabilities (Wainer, Feil-Seifer, Shell, & Mataric, 2007).

#### People perception

A natural reaction when we want to interact with someone is to direct our gaze towards the interested agent. The gaze feeds the communication, and conveys interest or attention to the interlocutor. It requires positioning the robot to make the most out of its sensors and to let the human talker know what the robot is actually paying attention to. Spontaneity during verbal communication involves two main behaviours, face and sound localization. Face localization is done applying OpenCV’s Haar feature-based cascade classifiers (Viola & Jones, 2001) to the images taken by the robot’s camera.

Sound source localization allows a robot to identify the direction of sound, and it is accomplished using microphone arrays, an algorithm based on TDOA (Time Difference of Arrival) approach (Bensky, 2016). The sound wave emitted by a source is received at slightly different times on each of the robot’s microphones, from the closest to the farthest. These differences are related to the current location of the emitting source. By using this relationship, the robot is able to retrieve the direction of the emitting source (azimuth and elevation angles) from the TDOAs measured on the different microphone pairs.

#### Key objects perception

The robot pays attention to different elements at different moments. The robot can be requested to reach the microphone to start its singing turn or it may need to go to rest to its chair. A colour tracking procedure enhanced with a Kalman Filter (Kalman, 1960) is used to produce a more robust behaviour against illumination conditions and balancing produced during walking. No location information in form of odometry or frame of reference is used because the location of those elements with respect to the robots varies depending on the scenario.

### Non-verbal communication

Body language represents the key to express feelings and helps people to understand sociability (Knight, 2011). While speaking or singing the poem, the robot has to generate credible body language that should shape and convey the information content.

#### Gestures

When the *bertsolari-s* are on the stage they are continuously conveying information, through facial expressions, body postures, movements or gestures, intentionally or not, about their emotional state. After identifying the main different states of the global behaviour, a gesture library composed by five different gesture sets have been defined to mimic verse-maker’s emotional behaviour on the stage (Rodriguez, Astigarraga, Ruiz, & Lazkano, 2016). At each state of the performance appropriate gesture set is selected. Three of the gesture sets represent states of the performance in which the robot does not receive any input from the environment (user or public).

- Thinking gestures: gestures that try to mimic human behavior while

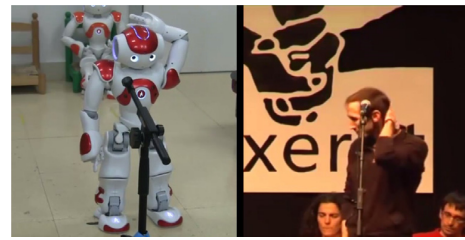


Fig. 6. Example of a thinking gesture.

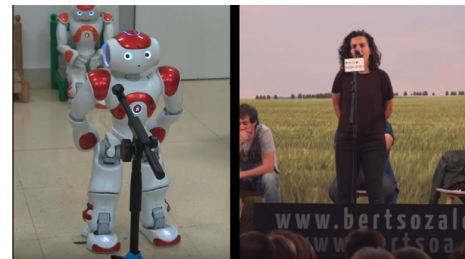


Fig. 7. Example of a singing gesture.

thinking (see Fig. 6).

- Singing gestures: those movements that verse-makers perform just after the improvisation process finishes and before they start singing (see Fig. 7).
- Waiting gestures: gestures that try to imitate human behavior while standing, such as change arms’ position and make movements with its head (see Fig. 8).  
In the remaining two gesture sets robot’s expression dynamically generated and adapted based on the sentiment of the text to be represented or the people’s reaction through applause feedback.
- Talking gestures: those movements that endow our humanoid robots with the ability to generate natural gesticulation movements enriched with other relevant social signals depending on sentiment processing. Head and arms movements, along with eye LED lighting and voice intonation are combined to make a humanoid robot express the sadness-happiness emotion continuum. Using sentiment analysis (Vicente, Saralegi, & Agerri, 2017) on the text that will be pronounced by the robot, we derive a value for emotion valence and coherently choose suitable parameters for the expressive elements. In this way, the robot has an adaptive expression generation during talking as it is shown in (Rodriguez, Martínez-Otzeta, Lazkano, & Ruiz, 2017). Fig. 9 shows some examples of generated gestures with emotion.
- Emotional reaction gestures: After the *bertsolari* sings a verse the audience responds applauding to express their opinion, and this reaction is reflected in the robot as an emotion gesture. As stated before, each applause feedback class has been represented with an emotion; in the next order Sad, Calm, Joy and Excited emotions correspond to Negative, Neutral, Positive and Very Positive applause classes (see Fig. 10).



Fig. 8. Example of a waiting gesture.

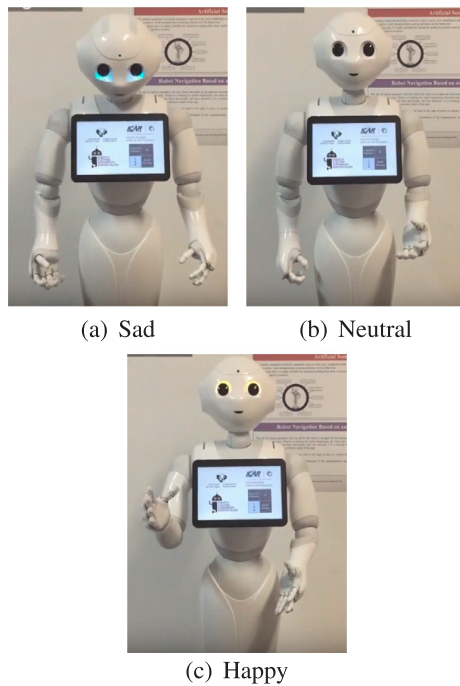


Fig. 9. Three basic emotions with their respective expressions.

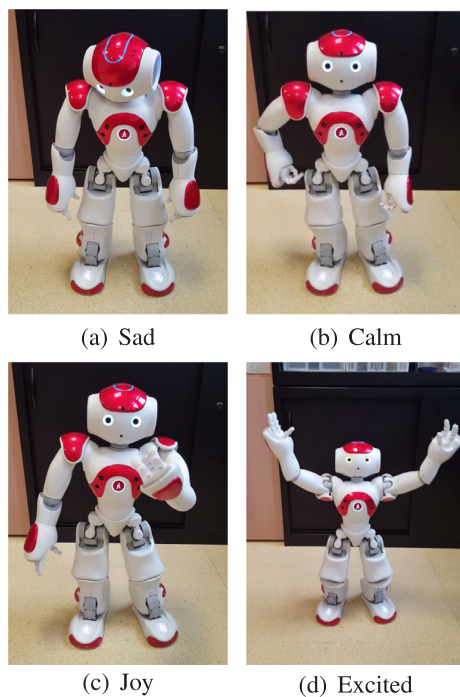


Fig. 10. Some examples of emotional reaction gestures.

### Evolution of the system through public performances

The robot's performance capabilities have been demonstrated in different events in a 5 years period. These public performances show the evolution of the BertsoBot project since its start up, when no humanoid platform was available, and up to now.

The objective of the live performances was not other than to bring

social robotics to the general public and, along with that, to receive audience's feedback about human-robot interaction.

First experiments were carried out by Tartalo and Galtxagorri, PeopleBot and Pioneer2DX mobile platforms from Omron Adept's MobileRobots.<sup>2</sup> Later on, we move to humanoid robots. We employed NAO and Pepper robots developed by Softbank Robotics.<sup>3</sup>

Table 2 shows the first and last version of the BertsoBot system, and their degree of implementation of the main cognitive functions. Fig. 11 depicts a functional description and inter-module communication of the final BertsoBot software architecture.

The "Performance Manager" is the behaviour that brings the coherence necessary to the system in order to follow the dynamic of a performance. It decides the action/actions the robot has to perform depending on the current state of the performance and the current body posture of the robot. The latter information is provided by the "Body Posture Awareness" module. The "People Perception" as well as the "Speech-Based Dialogue" behaviours allow the interaction with the emcee, while "Environmental Key Objects Perception" provides the robot with necessary skills to interact with environmental key objects. These interactions, usually executed as motion actions, are managed by the "Motion Control" behaviour. The verse is composed and sung by the "Verse Generation" process, and audience applauds, which affect the robot's emotional state, are captured and classified by "Feedback From the Audience" behaviour. The robot's emotional state is managed by the "Emotional Behaviour" module, which decides the emotional state that the robot must show considering both the feedback obtained from the audience and the emotion extracted from the text to be said by the robot. The robot body expression is managed by the "Adaptive Body Expression" behaviour, which generates appropriate gestures when talking, and chooses the predefined gestures to be applied from the appropriate gesture set at each state according to the state of the performance, the robot posture, and the emotional state of the robot.

- **2012/04: First public appearance:** Inauguration of the speaker's corner of our Campus. Paradoxically the most audacious one, due to the importance of the event and the preliminary state of the project. Tartalo and Galtxagorri were brought out and acted outdoor. No significant body language was shown, neither chatting was possible. Robots were mainly teleoperated and control software was Player/Stage.<sup>4</sup> Only the automatic verse generation system was embedded in wheeled robots. Video available.<sup>5</sup>
- **2013/05: Robots against bachelor students:** The robots took part in an event hold in the our faculty where they competed against some bertso-amateur students  
Tartalo was accompanied by NAO for the first time. Primary gestures were shown by NAO, that acted as the emcee semi-autonomously. NAO was controlled using Choregraphe, its native controller. Video available.<sup>6</sup>
- **2014/03: Women's day at the Faculty:** Our university annually celebrates the women's international day in a different center and in 2014 it was held at our Faculty. The program included a bertso event where two big professionals and two robots (NAO and Tartalo) took part.  
NAO showed improved chatting abilities, but still "unROSified". Primary gestures in NAO, that guided the event but semi-autonomously.<sup>7</sup>
- **2014/11: ScienceClub:** Club of Sciences events aim to disclose

<sup>2</sup> <http://www.adept.com/products/mobile-robots>.

<sup>3</sup> <https://www.ald.softbankrobotics.com/en/robots>.

<sup>4</sup> <http://playerstage.sourceforge.net/>.

<sup>5</sup> <https://www.youtube.com/watch?v=OpQBVmkzRWg>.

<sup>6</sup> <http://www.eitb.eus/eu/kultura/bertsolaritza/osoa/1350970/robot-bertsolariak-ixa-taldea-eta-ehuko-robotika-saila/>.

<sup>7</sup> <http://ehutb.ehu.es/es/video/index/uuid/531ec65f964be.html>.



**Table 2**  
Comparative table showing the cognitive abilities of the first (2012) and last (2017) version of the Bertso bot system.

	Galtxagorri/Tartalo (2012)	NAO/Pepper (2017)
Dialogue	Precompiled text	Basic chatting capabilities
Poetry Generation	Only one exercise, rhymes given: the system is given the four rhyming words and it is required to compose the bertso. Sentence-retrieval method	Two different exercises: rhymes given and topic given. Sentence-retrieval and n-gram generation methods. Affective state integrated in the creation process
Affective Perception	No	Audience applause as feedback
Interaction with the environment	Mainly teleoperated robots	Key object recognition integrated, fully autonomous robot
Body expression	Basic movements	Gesture sets to represent state of the performance. Automatically generated gesture sets based on the sentiment of the poem and people's reaction

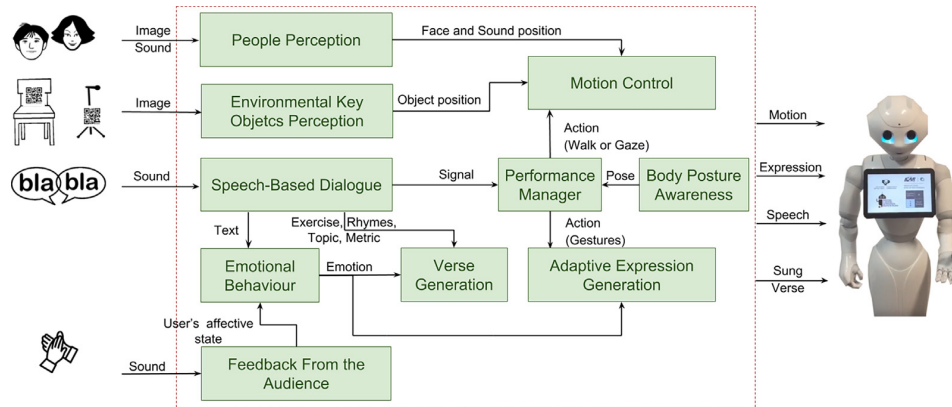


Fig. 11. Bertso bot's actual software framework.

science and technologies to the society. A dialogue with NAO entitled “Chatting with NAO” of approximately 10 min was presented. NAO acted alone and it was its first performance after being “ROSified”. However, it still acted semi-autonomously.

- **2015/11: ScienceClub:** Next year the title of the event was “NAO, an empathetic or just amusing robot?” Body gestures were integrated and chatting abilities were shown. The key object recognition was tested together with the face and sound localization behaviors. Video available.<sup>8</sup>
- **2016/02: Event at the Faculty:** A local event was organized at the Faculty in order to be able to evaluate the applause classification and emotional state gesture reproduction modules. Thinking and singing preamble gestures were used but there was no emcee neither environmental key objects to be easier for the audience to concentrate in the aspects that needed evaluation. Video available.<sup>9</sup>
- **2016/09: Closing of a Summer University Course:** Our university annually organizes several summer courses. Bertso bot was invited to the closing of a course entitled “Educational assessment: unresolved matter”. It was not a *bertso-saio* event but it covered all aspects of the interaction.
- **Lab demonstration** A rehearsal without audience recorded at our laboratory<sup>10</sup> exhibits the global behavior of the Bertso bot system in a performance similar to *bertsolari-s* events, in which two NAO act as verse-maker and the roll of the emcee is performed by *Galtxagorri*. The robotic emcee establishes the rules of the duel: who starts, the exercises and the flux of the performance.
- **2017-12: Lab demonstration** A local event realized in our laboratory in which Pepper robot acted alone as a verse-maker with a

human emcee guiding the event. Pepper composed poems on the fly and reacted according to the feedback received from the audience.

**Conclusions and further work**

We have presented in this paper an implementation of a software architecture designed for socially interacting robots. Although most of the implemented cognitive skills have been previously presented in other publications, this paper shows for the first time an overall integration of these components into a coherent and coordinated system for speech-based social HRI.

Specifically, we argue that oral improvised poetry can serve as a useful testbed for the development and evaluation of robots that interact with humans. The paper presents a speech-based humanoid poet-performer that can (1) listen to human commands and generate poems on demand; (2) perceive audiences feedback and react displaying the corresponding emotional response; and (3) generate natural gesticulation movements enriched with social signals depending on sentiment processing. We discuss each of the involved cognitive abilities, present working implementations and show how they are combined to achieve the fluent coordination and joint-action timing needed in live events.

This paper describes a dialogue system that combines basic chatting capabilities with a more elaborate oral poetry generation system. The proposed method not only generates novel poems, but also creates them conveying a certain attitude or state of mind. The system receives as an input the topic of the poem and the affective state (positive, neutral or negative) and tries to give as output a novel poem that satisfies formal constraints of rhyme and metric, shows coherent content related to the given topic and expresses them through the predetermined mood.

This work already allows robots to react and alter their behaviour during an event according to a specific audience’s natural feedback. Note that the developed system is not just a reactive system. The implemented on-line learning system allows the emotional system to adjust to the different audiences and, at the same time, to dynamically

<sup>8</sup> <https://www.youtube.com/watch?v=IMMXHWB2mZA>.

<sup>9</sup> <https://www.youtube.com/watch?v=SdxNgmV3CzA>.

<sup>10</sup> <https://www.youtube.com/watch?v=UNhvd2qbuaY>.

adapt to the overall state of the audience while the event progresses. Moreover, audience's feedback feeds the emotional reaction of the robot and also affects the poem generation system, adapting the text-sentiment to the perceived reaction. The adequateness of the overall system has been demonstrated through several real live performances.

Nevertheless, the proposed architecture has not achieved its final state and many improvements are under way. For instance, we are improving the dialogue capabilities of the robot, allowing an active learning for common ground knowledge acquisition, and translating our system to other experimental scenarios involving social interaction.

## Acknowledgments

This work has been partially supported by the Basque Government (IT900-16), the Spanish Ministry of Economy and Competitiveness MINECO/FEDER (TIN 201564395-R, MINECO/FEDER,EU). Author Rodriguez has received the UPV/EHU research Grant PIF13/104.

## References

- Agirrezabal, M., Alegria, I., Arrieta, B., & Hulden, M. (2012). BAD: An assistant tool for making verses in Basque. *Proceedings of the 6th workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 13–17). Association for Computational Linguistics.
- Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, 29(3), 313–341.
- Astigarraga, A., Jauregi, E., Lazkano, E., & Agirrezabal, M. (2014). Textual coherence in a verse-maker robot. *Human-computer systems interaction: backgrounds and applications: Vol. 3*, (pp. 275–287). Springer.
- Astigarraga, A., Martínez-Otzeta, J. M., Rodríguez, I., Sierra, B., & Lazkano, E. (2017). Emotional poetry generation. *International conference on speech and computer* (pp. 332–342). Springer.
- Augello, A., Cipolla, E., Infantino, I., Manfré, A., Pilato, G., & Vella, F. (2018). Social signs processing in a cognitive architecture for a humanoid robot. *Procedia Computer Science*, 123, 63–68.
- Augello, A., Infantino, I., Lieto, A., Pilato, G., Rizzo, R., & Vella, F. (2016). Artwork creation by a cognitive architecture integrating computational creativity and dual process approaches. *Biologically Inspired Cognitive Architectures*, 15, 74–86.
- Augello, A., Infantino, I., Pilato, G., Rizzo, R., & Vella, F. (2015). Creativity evaluation in a cognitive architecture. *Biologically Inspired Cognitive Architectures*, 11, 29–37.
- Bae, B.-C., Brunete, A., Malik, U., Dimara, E., Jermisurawong, J., & Mavridis, N., 2012. Towards an empathizing and adaptive storyteller system. In *Eighth artificial intelligence and interactive digital entertainment conference*.
- Bemelmans, R., Gelderblom, G. J., Jonker, P., & De Witte, L. (2012). Socially assistive robots in elderly care: A systematic review into effects and effectiveness. *Journal of the American Medical Directors Association*, 13(2), 114–120.
- Bensky, A. (2016). *Wireless positioning technologies and applications*. Artech House.
- Breazeal, C. L. (2004). *Designing sociable robots*. MIT Press.
- Bruce, A., Knight, J., Listopad, S., Magerko, B., & Nourbakhsh, I. R. (2000). Robot improv: Using drama to create believable agents. *Proceedings ICRA'00. IEEE international conference on robotics and automation, 2000: Vol. 4*, (pp. 4002–4008). IEEE.
- Cardoso, A., Veale, T., & Wiggins, G. A. (2009). Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI Magazine*, 30(3), 15.
- Chen, Y., Wu, F., Shuai, W., Wang, N., Chen, R., & Chen, X. (2015). Kejia robot—an attractive shopping mall guider. *International conference on social robotics* (pp. 145–154). Springer.
- Cherubini, A., Passama, R., Crosnier, A., Lasnier, A., & Fraise, P. (2016). Collaborative manufacturing with physical human–robot interaction. *Robotics and Computer-Integrated Manufacturing*, 40, 1–13.
- Chidambaram, V., Chiang, Y.-H., & Mutlu, B. (2012). Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. *Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction* (pp. 293–300). ACM.
- Choi, D., & Langley, P. (2018). Evolution of the ICARUS cognitive architecture. *Cognitive Systems Research*, 48, 25–38.
- Costa, S., Brunete, A., Bae, B.-C., & Mavridis, N. (2016). Emotional storytelling using virtual and robotic agents. *International Journal of Humanoid Robotics*, 15(03), 1850006.
- Dautenhahn, K., Woods, S., Kaouri, C., Walters, M. L., Koay, K. L., & Werry, I. (2005). What is a robot companion-friend, assistant or butler? *2005 IEEE/RSJ international conference on intelligent robots and systems, 2005. (IROS 2005)* (pp. 1192–1197). IEEE.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Fasola, J., & Mataric, M. J. (2012). Using socially assistive human–robot interaction to motivate physical exercise for older adults. *Proceedings of the IEEE*, 100(8), 2512–2526.
- Fernandez, J. M. A., & Bonarini, A. (2013). Theatrobot: A software architecture for a theatrical robot. *Conference towards autonomous robotic systems* (pp. 446–457). Springer.
- Fridin, M., & Belokopytov, M. (2014). Acceptance of socially assistive humanoid robot by preschool and elementary school teachers. *Computers in Human Behavior*, 33, 23–31.
- Garzia, J., Sarasua, J., & Egaña, A. (2001). *The art of bertsolaritza: Improvised basque verse singing*. Donostia, Spain: Bertsolari Liburuak. Bertsozaleak Kultur Elkarteak.
- Gervás, P. (2013). Computational modelling of poetry generation. *Artificial intelligence and poetry symposium, AISB convention 2013. The society for the study of artificial intelligence and the simulation of behaviour*. United Kingdom: University of Exeter.
- Gervás, P. (2015). Deconstructing computer poets: Making selected processes available as services. *Computational Intelligence*, 33(1), 3–31.
- Gómez Esteban, P., Cao, H.-L., De Beir, A., Van de Perre, G., Lefeber, D., & Vanderborght, B. (2016). A multilayer reactive system for robots interacting with children with autism. In *5th international symposium on new frontiers in human-robot interaction*. Sheffield, UK.
- Hernández, I., Navas, E., Murugarren, J. L., & Etxebarria, B. (2001). Description of the AhoTTS system for the Basque language. In *4th ISCA tutorial and research workshop (ITRW) on speech synthesis*.
- Heyer, C. (2010). Human-robot interaction and future industrial robotics applications. *2010 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 4749–4754). IEEE.
- Hoffman, G., 2011. On stage: Robots as performers. In *RSS 2011 workshop on human-robot interaction: Perspectives and contributions to robotics from the human sciences* (Vol. 1). Los Angeles, CA.
- Kachouie, R., Sedighadel, S., Khosla, R., & Chu, M.-T. (2014). Socially assistive robots in elderly care: a mixed-method systematic literature review. *International Journal of Human-Computer Interaction*, 30(5), 369–393.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, 82(Series D), 35–45.
- Kanda, A., Arai, M., Suzuki, R., Kobayashi, Y., & Kuno, Y. (2014). Recognizing groups of visitors for a robot museum guide tour. *2014 7th international conference on human system interactions (HSI)* (pp. 123–128). IEEE.
- Kanda, T., Shimada, M., & Koizumi, S. (2012). Children learning with a social robot. *Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction* (pp. 351–358). ACM.
- Knight, H. (2011). Eight lessons learned about non-verbal interactions through robot theater. *Social Robotics*, 42–51.
- Kosuge, K., Hayashi, T., Hirata, Y., & Tobiyama, R. (2003). Dance partner robot – Ms DanceR. *Proceedings 2003 IEEE/RSJ international conference on intelligent robots and systems, 2003. (IROS 2003): Vol. 4*, (pp. 3459–3464). IEEE.
- Kraemer, F., Rodríguez, I., Parra, O., Ruiz, T., & Lazkano, E. (2016). Minstrel robots: Body language expression through applause evaluation. *2016 IEEE-RAS 16th international conference on humanoid robots (humanoids)* (pp. 332–337). IEEE.
- Laird, J. E., Kinkade, K. R., Mohan, S., & Xu, J. Z. (2012). Cognitive robotics using the Soar cognitive architecture. *Cognitive Robotics AAAI Technical Report, WS-12*, 6, 46–54.
- Lamb, C., Brown, D. G., & Clarke, C. L. (2016). A taxonomy of generative poetry techniques. In *Bridges finland conference proceedings* (pp. 195–202).
- Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2), 141–160.
- Lee, S. A., & Liang, Y. (2016). The role of reciprocity in verbally persuasive robots. *Cyberpsychology, Behavior, and Social Networking*, 19(8), 524–527.
- Lord, A. B., Mitchell, S. A., & Nagy, G. (2000). *The singer of tales. Vol. 24 of Harvard studies in comparative literature*. Harvard University Press.
- Luria, M., Hoffman, G., Megidish, B., Zuckerman, O., & Park, S. (2016). Designing Vyo, a robotic smart home assistant: Bridging the gap between device and social agent. *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 1019–1025). IEEE.
- Moyle, W., Cooke, M., Beattie, E., Jones, C., Klein, B., Cook, G., Gray, C., et al. (2013). Exploring the effect of companion robots on emotional expression in older adults with dementia: A pilot randomized controlled trial. *Journal of Gerontological Nursing*, 39(5), 46–53.
- Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.
- Oliveira, H. G. (2017). A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th international conference on natural language generation* (pp. 11–20).
- Pihel, E. (1996). A purified freestyle: Homer and hip hop. *Oral Tradition*, 11(2), 249–269.
- Rashed, M. G., Suzuki, R., Lam, A., Kobayashi, Y., & Kuno, Y. (2015). Toward museum guide robots proactively initiating interaction with humans. *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts* (pp. 1–2). ACM.
- Rodríguez, I., Astigarraga, A., Ruiz, T., & Lazkano, E. (2016). Singing minstrel robots, a means for improving social behaviors. *2016 IEEE international conference on robotics and automation (ICRA)* (pp. 2902–2907). IEEE.
- Rodríguez, I., Martínez-Otzeta, J. M., Lazkano, E., & Ruiz, T. (2017). Adaptive emotional chatting behavior to increase the sociability of robots. *International conference on social robotics* (pp. 666–675). Springer.
- Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. *Cognition and Multi-agent Interaction*, 79–99.
- Tapus, A., Tapus, C., & Mataric, M. J. (2009). The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. *IEEE international conference on rehabilitation robotics, 2009. ICORR 2009* (pp. 924–929). IEEE.
- Thórisson, K., & Helgasson, H. (2012). Cognitive architectures and autonomy: A comparative review. *Journal of Artificial General Intelligence*, 3(2), 1–30.
- Vernon, D., Metta, G., & Sandini, G. (2007a). The iCub cognitive architecture: Interactive development in a humanoid robot. *IEEE 6th international conference on development and learning, 2007. ICDL 2007* (pp. 122–127). IEEE.

- Vernon, D., Metta, G., & Sandini, G. (2007b). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation*, 11(2), 151–180.
- Vernon, D., von Hofsten, C., & Fadiga, L. (2016). Desiderata for developmental cognitive architectures. *Biologically Inspired Cognitive Architectures*, 18, 116–127.
- Vicente, I. S., Saralegi, X., & Agerri, R. (2017). EliXa: A modular and flexible ABSA platform. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 748–752).
- Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition, 2001. CVPR 2001. Vol. 1. Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition, 2001. CVPR 2001* (pp. 511–518). IEEE.
- Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. *The 16th IEEE international symposium on robot and human interactive communication, 2007. RO-MAN 2007* (pp. 872–877). IEEE.
- Wisspeintner, T., Van Der Zant, T., Iocchi, L., & Schiffer, S. (2009). RoboCup@Home: Scientific competition and benchmarking for domestic service robots. *Interaction Studies*, 10(3), 392–426.
- Wu, Y., Wang, R., Tay, Y. L., & Wong, C. J. (2017). Investigation on the roles of human and robot in collaborative storytelling. *Asia-pacific signal and information processing association annual summit and conference (APSIPA ASC), 2017* (pp. 063–068). IEEE.

# Bibliography

- [1] M. Agirrezabal, I. Alegria, B. Arrieta, and M. Hulden. BAD: An assistant tool for making verses in Basque. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 13–17. Association for Computational Linguistics, 2012.
- [2] D. W. Aha, D. Kibler, and M. K. Albert. *Machine learning*, chapter Instance-based learning algorithms, pages 37–66. Springer, 1991.
- [3] I. Almetwally and M. Mallem. Real-time tele-operation and tele-walking of humanoid robot nao using kinect depth camera. In *International Conference on Networking, Sensing and Control (ICNSC)*, pages 463–466. IEEE, 2013.
- [4] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu. Conversational gaze aversion for humanlike robots. In *International Conference on Human-Robot Interaction (HRI)*, pages 25–32. ACM/IEEE, 2014.
- [5] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.
- [6] A. Astigarraga. *Bertsobot: Gizaki-Robot Arteko Komunikazio eta Elkarrekintzarako Portaerak*. PhD thesis, UPV/EHU, 2017.
- [7] A. Astigarraga, M. Agirrezabal, E. Lazkano, E. Jauregi, and B. Sierra. Bertsobot: the first minstrel robot. In *International Conference on Human System Interaction (HSI)*, pages 129–136. IEEE, 2013.
- [8] A. Astigarraga, E. Jauregi, E. Lazkano, and M. Agirrezabal. Textual coherence in a verse-maker robot. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*, pages 275–287. Springer, 2014.
- [9] A. Astigarraga, E. Lazkano, I. Ranó, B. Sierra, and I. Zarautz. Sorgin: a software framework for behavior control implementation. In *International Conference on Control Systems and Computer Science (CSCS)*, pages 243–248, 2003.
- [10] A. Astigarraga, J. M. Martínez-Otzeta, I. Rodriguez, B. Sierra, and E. Lazkano. Emotional poetry generation. In *International Conference on Speech and Computer*, pages 332–342. Springer, 2017.

- 
- [11] A. Augello, E. Cipolla, I. Infantino, A. Manfrè, G. Pilato, and F. Vella. Creative Robot Dance with Variational Encoder. In *International Conference on Computational Creativity*, 2017.
- [12] A. Augello, I. Infantino, A. Manfrè, G. Pilato, F. Vella, and A. Chella. Creation and cognition for humanoid live dancing. *Robotics and Autonomous Systems*, 86:128–137, 2016.
- [13] A. Augello, I. Infantino, U. Maniscalco, G. Pilato, and F. Vella. Robot inner perception capability through a soft somatosensory system. *International Journal of Semantic Computing*, 12(01):59–87, 2018.
- [14] G. H. Ballantyne and F. Moll. The da Vinci telerobotic surgical system: the virtual operative field and telepresence surgery. *Surgical Clinics*, 83(6):1293–1304, 2003.
- [15] T. Bänziger and K. R. Scherer. The role of intonation in emotional expressions. *Speech communication*, 46(3):252–267, 2005.
- [16] E. I. Barakova and T. Lourens. Expressing and interpreting emotional movements in social games with robots. *Personal and ubiquitous computing*, 14(5):457–467, 2010.
- [17] A. Beck, L. Cañamero, A. Hiole, L. Damiano, P. Cosi, F. Tesser, and G. Sommavilla. Interpretation of emotional body language displayed by a humanoid robot: a case study with children. *International Journal of Social Robotics*, 5(3):325–334, 2013.
- [18] A. Beck, Z. Yumak, and N. Magnenat-Thalmann. Body movements generation for virtual characters and social robots. In *Social signal processing*, chapter 20, pages 273–286. Cambridge University Press, 2017.
- [19] J. M. Beer, A. D. Fisk, and W. A. Rogers. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction*, 3(2):74, 2014.
- [20] G. A. Bekey. *Autonomous robots: from biological inspiration to implementation and control*. MIT press, 2005.
- [21] R. Bemelmans, G. J. Gelderblom, P. Jonker, and L. De Witte. Socially assistive robots in elderly care: A systematic review into effects and effectiveness. *Journal of the American Medical Directors Association*, 13(2):114–120, 2012.
- [22] A. Bensky. *Wireless positioning technologies and applications*. Artech House, 2016.
- [23] J. J. Biesiadecki, E. T. Baumgartner, R. G. Bonitz, B. Cooper, F. R. Hartman, P. C. Leger, M. W. Maimone, S. A. Maxwell, A. Trebi-Ollennu, E. W. Tunstel, et al. Mars exploration rover surface operations: Driving opportunity at meridiani planum. *IEEE robotics & automation magazine*, 13(2):63–71, 2006.

- [24] Blue Frog Robotics. Buddy. <http://www.bluefrogrobotics.com/en/buddy/>, [accessed May 9, 2018].
- [25] P. Boissy, S. Brière, H. Corriveau, A. Grant, M. Lauria, and F. Michaud. Usability testing of a mobile robotic system for in-home telerehabilitation. In *International Conference on Engineering in Medicine and Biology Society (EMBC)*, pages 1839–1842. IEEE, 2011.
- [26] C. Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2):119–155, 2003.
- [27] C. Breazeal. *Designing sociable robots*. MIT press, 2004.
- [28] C. Breazeal, M. Siegel, M. Berlin, J. Gray, R. Grupen, P. Deegan, J. Weber, K. Narendran, and J. McBean. Mobile, dexterous, social robots for mobile manipulation and human-robot interaction. In *International Conference and Exhibition on Computer Graphics and Interactive Techniques*, page 27. ACM, 2008.
- [29] M. Bretan, G. Hoffman, and G. Weinberg. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies*, 78:1–16, 2015.
- [30] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai. Highlight sound effects detection in audio stream. In *International Conference on Multimedia and Expo (ICME)*, volume 3, pages III–37. IEEE, 2003.
- [31] Cambridge Medical Robotics (CMR) Surgical. Versius robot. <https://cmrsurgical.com/versius/>, [accessed May 9, 2018].
- [32] L. Cañamero and M. Lewis. Making new “new AI” friends: designing a social robot for diabetic children from an embodied AI perspective. *International Journal of Social Robotics*, 8(4):523–537, 2016.
- [33] Á. Castro-González, M. Malfaz, J. F. Gorostiza, and M. A. Salichs. Learning behaviors by an autonomous social robot with motivations. *Cybernetics and Systems*, 45(7):568–598, 2014.
- [34] A. Chatley, K. Dautenhahn, M. W. abd D.S. Syrdal, and B. Christianson. Theater as a discussion tool in human-robot interaction experiments. a pilot study. In *International Conference on Advances in Computer-Human Interactions (ACHI)*, pages 73–78. IEEE, 2010.
- [35] G. Cicirelli, C. Attolico, C. Guaragnella, and T. D’Orazio. A kinect-based gesture recognition approach for a natural human robot interface. *International Journal of Advanced Robotic Systems*, 12(3):22, 2015.

- [36] J.-A. Claret, G. Venture, and L. Basañez. Exploiting the robot kinematic redundancy for emotion conveyance to humans as a lower priority task. *International Journal of Social Robotics*, 9(2):277–292, 2017.
- [37] A. Clark. An embodied cognitive science? *Trends in Cognitive Sciences*, 3(9):345–351, 1999.
- [38] M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, 2004.
- [39] J. Crumpton and C. L. Bethel. A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics*, 8(2):271–285, 2016.
- [40] N. Dael, M. Mortillaro, and K. R. Scherer. The body action and posture coding system (bap): Development and reliability. *Journal of Nonverbal Behavior*, 36(2):97–121, 2012.
- [41] K. Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704, 2007.
- [42] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
- [43] A. Deshmukh, B. Craenen, A. Vinciarelli, and M. E. Foster. Modulating the non-verbal social signals of a humanoid robot. In *International Conference on Multimodal Interaction (ICMI)*, pages 508–509. ACM, 2017.
- [44] J. Dias, S. Mascarenhas, and A. Paiva. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Emotion modeling*, pages 44–56. Springer, 2014.
- [45] M. Diftler, T. Ahlstrom, R. Ambrose, N. Radford, C. Joyce, N. De La Pena, A. Parsons, and A. Noblitt. Robonaut 2 - Initial activities on-board the ISS. In *Aerospace Conference*, pages 1–12. IEEE, 2012.
- [46] G. Du, P. Zhang, J. Mai, and Z. Li. Markerless kinect-based hand tracking for robot teleoperation. *International Journal of Advanced Robotic Systems*, 9(2):36–45, 2012.
- [47] B. R. Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3-4):177–190, 2003.
- [48] B. J. Dunstan, D. Silvera-Tawil, J. T. Koh, and M. Velonaki. Cultural robotics: Robots as participants and creators of culture. In *International Workshop in Cultural Robotics*, pages 3–13. Springer, 2015.

- [49] P. Ekman. Are there basic emotions? *Psychological Review*, pages 550–553, 1992.
- [50] P. Ekman and W. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. *CA: Consulting Psychologists Press*, 1978.
- [51] J. M. G. Enrique Castillo and A. S. Hadi. *Learning Bayesian Networks. Expert Systems and Probabilistic Network Models*. Monographs in computer science. New York: Springer-Verlag, 1997.
- [52] M. S. Erden. Emotional postures for the humanoid-robot NAO. *International Journal of Social Robotics*, 5(4):441–456, 2013.
- [53] B. Everitt and D. Hand. *Finite mixture distributions*. Chapman and Hall, 1981.
- [54] F. Faber, M. Bennewitz, C. Eppner, A. Gorog, C. Gonsior, D. Joho, M. Schreiber, and S. Behnke. The humanoid museum tour guide robotinho. In *International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 891–896. IEEE, 2009.
- [55] D. Feil-Seifer and M. J. Matarić. Defining socially assistive robotics. In *International Conference on Rehabilitation Robotics (ICORR)*, pages 465–468. IEEE, 2005.
- [56] J. Feldmaier, T. Marmat, J. Kuhn, and K. Diepold. Evaluation of a RGB-LED-based Emotion Display for Affective Agents. *arXiv preprint arXiv:1612.07303*, 2016.
- [57] J. Fernandez and A. Bonarini. TheatreBot: A Software Architecture for a Theatrical Robot. In *Towards Autonomous Robotic Systems (TAROS)*, pages 446–457. Springer, Birmingham, UK, 2014.
- [58] T. Fischer, J.-Y. Puigbò, D. Camilleri, P. D. Nguyen, C. Moulin-Frier, S. Lalée, G. Metta, T. J. Prescott, Y. Demiris, and P. F. Verschure. iCub-HRI: A Software Framework for Complex Human–Robot Interaction Scenarios on the iCub Humanoid Robot. *Frontiers in Robotics and AI*, 5:22–30, 2018.
- [59] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3-4):143–166, 2003.
- [60] T. Fong and C. Thorpe. Vehicle teleoperation interfaces. *Autonomous robots*, 11(1):9–18, 2001.
- [61] J. Forlizzi and C. DiSalvo. Service robots in the domestic environment: a study of the roomba vacuum in the home. In *SIGCHI/SIGART conference on Human-robot interaction*, pages 258–265. ACM, 2006.



- [62] M. E. Foster, M. Giuliani, A. Isard, C. Matheson, J. Oberlander, and A. Knoll. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1818–1823, 2009.
- [63] N. Fraser. Assessment of interactive systems. In *Handbook of standards and resources for spoken language systems*, pages 564–615. Mouton de Gruyter, 1998.
- [64] M. Fridin. Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Computers & education*, 70:53–64, 2014.
- [65] M. Fridin and M. Belokopytov. Acceptance of socially assistive humanoid robot by preschool and elementary school teachers. *Computers in Human Behavior*, 33:23–31, 2014.
- [66] L. Gallardo-Estrella and A. Poncela. Human/robot interface for voice teleoperation of a robotic platform. In *International Work-Conference on Artificial Neural Networks*, pages 240–247. Springer, 2011.
- [67] I. Goodfellow. NIPS Tutorial: Generative Adversarial Networks. *ArXiv e-prints*, dec 2017.
- [68] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
- [69] A. Graves. Generating sequences with recurrent neural networks. Technical report, Cornell University, 2013.
- [70] H.-M. Gross, H. Boehme, C. Schroeter, S. Müller, A. König, E. Einhorn, C. Martin, M. Merten, and A. Bley. TOOMAS: interactive shopping guide robots in everyday use-final implementation and experiences from long-term field trials. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 2005–2012. IEEE, 2009.
- [71] E. Guizzo. Japan earthquake: Robots help search for survivors. <https://spectrum.ieee.org/automaton/robotics/industrial-robots/japan-earthquake-robots-help-search-for-survivors>, [accessed May 9, 2018].
- [72] A. Gupta, S. Savarese, A. Alahi, et al. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.
- [73] J. W. Hart and B. Scassellati. Mirror perspective-taking with a humanoid robot. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 1990–1996, 2012.

- [74] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural Networks for Perception*, pages 65–93. Elsevier, 1992.
- [75] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer. Theory of Mind (ToM) on robots: a functional neuroimaging study. In *International Conference on Human-Robot Interaction*, pages 335–342. ACM/IEEE, 2008.
- [76] I. Hernáez, E. Navas, J. L. Murugarren, and B. Etxebarria. Description of the AhoTTS system for the Basque language. In *ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.
- [77] K. Höök. Affective loop experiences: designing for interactional embodiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:3585–3595, 2009.
- [78] A. Hristoskova, C. Agüero, M. Veloso, and F. Turck. Personalized guided tour by multiple robots through semantic profile definition and dynamic redistribution of participants. In *International Cognitive Robotics Workshop at AAAI, Toronto, Canada*, volume 1, pages 39–45, 2012.
- [79] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *International AAAI Conference on Weblogs and Social Media*, pages 216–225, 2014.
- [80] I. Infantino. Affective human-humanoid interaction through cognitive architecture. In *The Future of Humanoid Robots-Research and Applications*, pages 147–164. InTech, 2012.
- [81] International Federation of Robotics. (ifr). <http://www.ifr.org/service-robots/>, [accessed May 9, 2018].
- [82] M. Jang, J. Kim, and B.-K. Ahn. A software framework design for social human-robot interaction. In *International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 411–412. IEEE, 2015.
- [83] E. Jauregi. *Advances towards behaviour-based indoor robotic exploration*. PhD thesis, UPV/EHU, 2011.
- [84] Jibo Inc. Jibo. <https://www.jibo.com/>, [accessed May 9, 2018].
- [85] E. Jochum, E. Vlachos, A. Christoffersen, S. Gridsted Nielsen, I. A. Hameed, and Z. Tan. Using Theater to Study Interaction with Care Robots. *International Journal of Social Robotics*, 8:457–470, 2016.
- [86] D. O. Johnson, R. H. Cuijpers, and D. van der Pol. Imitating human emotions with artificial facial expressions. *International Journal of Social Robotics*, 5(4):503–513, 2013.

- [87] T. Kanda, M. Shimada, and S. Koizumi. Children learning with a social robot. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 351–358. ACM, 2012.
- [88] K. Katevas, P. G. Healey, and M. T. Harris. Robot stand-up: engineering a comic performance. In *International Conference on Humanoid Robots (Humanoids)*, 2014.
- [89] S. J. Kim, Y. Jeong, S. Park, K. Ryu, and G. Oh. A Survey of Drone use for Entertainment and AVR (Augmented and Virtual Reality). In *Augmented Reality and Virtual Reality*, pages 339–352. Springer, 2018.
- [90] H. Knight. Eight lessons learned about non-verbal interactions through robot theater. In *International Conference on Social Robotics (ICSR)*, pages 42–51. Springer, 2011.
- [91] H. Knight, S. Satkin, V. Ramakrishna, and S. Divvala. A savvy robot standup comic: Online learning through audience tracking. In *International Conference on Tangible, Embedded and Embodied Interaction (TEI)*, 2011.
- [92] S. Koceski and N. Koceska. Evaluation of an assistive telepresence robot for elderly healthcare. *Journal of Medical Systems*, 40(5):121–127, 2016.
- [93] J. Koenemann, F. Burget, and M. Bennewitz. Real-time imitation of human whole-body motions by humanoids. In *International Conference on Robotics and Automation (ICRA)*, pages 2806–2812. IEEE, 2014.
- [94] J. Kwon and F. C. Park. Using Hidden Markov Models to Generate Natural Humanoid Movement. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2006.
- [95] H. Lagercrantz and J.-P. Changeux. The emergence of human consciousness: from fetal to neonatal life. *Pediatric Research*, 65(3):255, 2009.
- [96] I. Lane, V. Prasad, G. Sinha, A. Umuhoza, S. Luo, A. Chandrashekar, and A. Raux. HRItk: the human-robot interaction ToolKit rapid development of speech-centric interactive systems in ROS. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 41–44. Association for Computational Linguistics, 2012.
- [97] E. Lazkano. *Pautas para el desarrollo incremental de una arquitectura de control basada en el comportamiento para la navegación de robots en entornos semi-estructurados*. PhD thesis, UPV/EHU, 2004.
- [98] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

- [99] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.
- [100] M. Lhommet and S. Marsella. Expressing emotion through posture and gesture. In *The Oxford Handbook of Affective Computing*, pages 273–285. Oxford University Press, 2015.
- [101] B. Li. Robot. <http://www.blancali.com/en/event/99/robot>, [accessed May 31, 2018].
- [102] C. Li, C. Yang, P. Liang, A. Cangelosi, and J. Wan. Development of kinect based teleoperation of NAO robot. In *International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 133–138. IEEE, 2016.
- [103] C. Lin, C. Tseng, W. Teng, W. Lee, C. Kuo, H. Gu, K. Chung, and C. Fahn. The realization of robot theater: Humanoid robots and theatric performance. In *International Conference on Advanced Robotics (ICAR)*, pages 1–6, Munich, Germany, 2009. IEEE.
- [104] J. López, D. Pérez, M. Santos, and M. Cacho. GuideBot. A tour guide system based on mobile robots. *International Journal of Advanced Robotic Systems*, 10(11):381–394, 2013.
- [105] J. López, D. Pérez, E. Zalama, and J. Gómez-García-Bermejo. Bellbot-a hotel assistant system using mobile robots. *International Journal of Advanced Robotic Systems*, 10(1):40–50, 2013.
- [106] D. Lu. Ontology of Robot Theatre. In *ICRA Workshop on Robotics and Performing Arts: Reciprocal Influences*, 2012.
- [107] K. F. MacDorman and H. Ishiguro. The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3):297–337, 2006.
- [108] A. Manfrè, I. Infantino, F. Vella, and S. Gaglio. An automatic system for humanoid dance creation. *Biologically Inspired Cognitive Architectures*, 15:1–9, 2016.
- [109] E. Marder-Eppstein, E. Berger, T. Foote, B. Gerkey, and K. Konolige. The office marathon: Robust navigation in an indoor office environment. In *International Conference on Robotics and Automation (ICRA)*, pages 300–307. IEEE, 2010.
- [110] M. J. Matarić. *The robotics primer*. MIT Press, 2007.

- [111] D. Matsui, T. Minato, K. F. MacDorman, and H. Ishiguro. Generating natural motion in an android by mapping human motion. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 3301–3308. IEEE/RSJ, 2005.
- [112] S. McGlynn, B. Snook, S. Kemple, T. L. Mitzner, and W. A. Rogers. Therapeutic robots for older adults: investigating the potential of paro. In *International Conference on Human-Robot Interaction*, pages 246–247. ACM/IEEE, 2014.
- [113] D. McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.
- [114] M. F. McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, 34(1):90–169, 2002.
- [115] P. Michel, K. Gold, and B. Scassellati. Motion-based robotic self-recognition. In *International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2763–2768. IEEE, 2004.
- [116] R. Murphy. Activities of the Rescue Robots at the World Trade Center from 11-21 September 2001. *IEEE Robotics & Automation Magazine*, pages 50–61, 2004.
- [117] R. Murphy, D. Shell, A. Guerin, B. Duncan, B. Fine, K. Pratt, and T. Zournetos. A Midsummer Night’s Dream (with flying robots). *Autonomous Robots*, 30(2):143–156, 2011.
- [118] K. Ogawa, K. Taura, and H. Ishiguro. Possibilities of Androids as Poetry-reciting Agent. In *International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 565–570. IEEE, September 2012.
- [119] Y. Ou, J. Hu, Z. Wang, Y. Fu, X. Wu, and X. Li. A real-time human imitation system using kinect. *International Journal of Social Robotics*, 7(5):587–600, 2015.
- [120] A. Paiva, I. Leite, and T. Ribeiro. Emotion modelling for social robots. In *The Oxford Handbook of Affective Computing*, chapter 21, pages 296–308. Oxford University Press, 2015.
- [121] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.
- [122] R. Pérula-Martínez. *Autonomous decision-making for socially interactive robots*. PhD thesis, Universidad Carlos III de Madrid. Departamento de Ingeniería de Sistemas y Automática, 2017.
- [123] R. Pfeifer and C. Scheier. *Understanding Intelligence*. MIT press, 2001.

- [124] S. Pieska, M. Luimula, J. Jauhiainen, and V. Spiz. Social service robots in wellness and restaurant applications. *Journal of Communication and Computer*, 10(1):116–123, 2013.
- [125] M. J. Powell, H. Zhao, and A. D. Ames. Motion primitives for human-inspired bipedal robotic locomotion: walking and stair climbing. In *International Conference on Robotics and Automation (ICRA)*, pages 543–549. IEEE, 2012.
- [126] J. Preece, Y. Rogers, and H. Sharp. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons, 2015.
- [127] A. Pronobis and R. P. N. Rao. Learning Deep Generative Spatial Models for Mobile Robots. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 755–762, 2017.
- [128] J. R. Quinlan. C4.5: Programming for machine learning. *Morgan Kaufmann*, 38:48, 1993.
- [129] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.
- [130] P.-L. P. Rau, Y. Li, and J. Liu. Effects of a social robot’s autonomy and group orientation on human decision-making. *Advances in Human-Computer Interaction*, 2013:1–13, 2013.
- [131] V. R. Reddy and T. Chattopadhyay. Human activity recognition from kinect captured data using stick model. In *International Conference on Human-Computer Interaction (HCI)*, pages 305–315. Springer, 2014.
- [132] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, volume 48, pages 1060–1069. PMLR, 2016.
- [133] H. Robotics. Sophia. <http://www.hansonrobotics.com/robot/sophia/>, [accessed June 17, 2018].
- [134] S. Rosenthal, J. Biswas, and M. Veloso. An effective personal mobile robot agent through symbiotic human-robot interaction. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, volume 1, pages 915–922. International Foundation for Autonomous Agents and Multiagent Systems, 2010.
- [135] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.
- [136] D. Sakamoto, T. Kanda, T. Ono, H. Ishiguro, and N. Hagita. Android as a telecommunication medium with a human-like presence. In *International Conference on Human-Robot Interaction (HRI)*, pages 193–200. ACM/IEEE, 2007.

- [137] I. San Vicente, X. Saralegi, and R. Agerri. EliXa: A modular and flexible ABSA platform. In *International Workshop on Semantic Evaluation (SemEval)*, pages 748–752, 2015.
- [138] M. Sarabia, R. Ros, and Y. Demiris. Towards an open-source social middleware for humanoid robots. In *International Conference on Humanoid Robots (Humanoids)*, pages 670–675. IEEE, 2011.
- [139] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, and G. K. Santhanam. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. *Monthly Notices of the Royal Astronomical Society: Letters*, 467(1):L110–L114, 2017.
- [140] A. Setapen, M. Quinlan, and P. Stone. Beyond teleoperation: Exploiting human motor skills with marionet. In *AAMAS Workshop on Agents Learning Interactively from Human Teachers (ALIHT)*, 2010.
- [141] C. Shi, S. Satake, T. Kanda, and H. Ishiguro. A Robot that Distributes Flyers to Pedestrians in a Shopping Mall. *International Journal of Social Robotics*, pages 1–17, 2017.
- [142] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita. Interactive humanoid robots for a science museum. In *SIGCHI/SIGART Conference on Human-Robot Interaction (HRI)*, pages 305–312. ACM, 2006.
- [143] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.
- [144] N. E. Sian, K. Yokoi, S. Kajita, F. Kanehiro, and K. Tanie. Whole body teleoperation of a humanoid robot development of a simple master device using joysticks. *Journal of the Robotics Society of Japan*, 22(4):519–527, 2004.
- [145] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza. *Introduction to autonomous mobile robots. Second Edition*. MIT press, 2011.
- [146] H. Song, D. Kim, M. Park, and J. Park. Tele-operation between human and robot arm using wearable electronic device. *IFAC Proceedings Volumes*, 41(2):2430–2435, 2008.
- [147] W. Song, X. Guo, F. Jiang, S. Yang, G. Jiang, and Y. Shi. Teleoperation humanoid robot control system based on kinect sensor. In *International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 2, pages 264–267. IEEE, 2012.
- [148] S. W. Squyres, R. E. Arvidson, J. F. Bell, J. Brückner, N. A. Cabrol, W. Calvin, M. H. Carr, P. R. Christensen, B. C. Clark, L. Crumpler, et al. The Spirit rover’s Athena science investigation at Gusev crater, Mars. *Science*, 305(5685):794–799, 2004.

- [149] L. Susperregi, I. Fernandez, A. Fernandez, S. Fernandez, I. Maurtua, and I. L. de Vallejo. Interacting with a robot: a guide robot understanding natural language instructions. In *International Conference on Ubiquitous Computing and Ambient Intelligence (UCAMI)*, pages 185–192. Springer, 2012.
- [150] A. Taheri, A. Meghdari, M. Alemi, H. Pouretamad, P. Poorgoldooz, and M. Roohbakhsh. Social robots and teaching music to autistic children: Myth or reality? In *International Conference on Social Robotics (ICSR)*, pages 541–550. Springer, 2016.
- [151] H. Takanobu, H. Tabayashi, S. Narita, A. Takanishi, E. Guglielmelli, and P. Dario. Remote interaction between human and humanoid robot. *Journal of Intelligent and Robotic Systems*, 25(4):371–385, 1999.
- [152] F. Tanaka, T. Takahashi, S. Matsuzoe, N. Tazawa, and M. Morita. Telepresence robot helps children in communicating with teachers who speak a different language. In *International Conference on Human-Robot Interaction (HRI)*, pages 399–406. ACM/IEEE, 2014.
- [153] A. K. Tanwani. *Generative Models for Learning Robot Manipulation*. PhD thesis, École Polytechnique Fédérale de Laussane (EPFL), 2018.
- [154] R. Y. Tara, P. I. Santosa, and T. B. Adji. Sign language recognition in robot teleoperation using centroid distance fourier descriptors. *International Journal of Computer Applications*, 48(2):8–12, 2012.
- [155] S. Thrun. Toward a framework for human-robot interaction. *Human-Computer Interaction*, 19(1):9–24, 2004.
- [156] S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, et al. MINERVA: A second-generation museum tour-guide robot. In *International Conference on Robotics and automation (ICRA)*, volume 3, pages 1999–2005. IEEE, 1999.
- [157] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT press, 2005.
- [158] M. Tielman, M. Neerinx, J.-J. Meyer, and R. Looije. Adaptive emotional expression in robot-child interaction. In *International Conference on Human-Robot Interaction (HRI)*, pages 407–414. ACM/IEEE, 2014.
- [159] P. Trahanias, W. Burgard, A. Argyros, D. Hahnel, H. Baltzakis, P. Pfaff, and C. Stachniss. TOURBOT and WebFAIR: Web-operated mobile robots for telepresence in populated exhibitions. *IEEE Robotics & Automation Magazine*, 12(2):77–89, 2005.
- [160] D. Troniak, J. Sattar, A. Gupta, J. J. Little, W. Chan, E. Caliskan, E. Croft, and M. Van der Loos. Charlie Rides the Elevator—Integrating Vision, Navigation and Manipulation towards Multi-floor Robot Locomotion. In *International Conference on Computer and Robot Vision (CRV)*, pages 1–8. IEEE, 2013.



- [161] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A Generative Model for Raw Audio. In *ISCA Speech Synthesis Workshop*, pages 755–762, 2016.
- [162] A. Van Den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. C. Rus, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In *International Conference on Machine Learning (ICML)*, 2018.
- [163] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *International Conference on Human-Robot Interaction (HRI)*, pages 42–52. ACM/IEEE, 2017.
- [164] M. Veloso, J. Biswas, B. Coltin, and S. Rosenthal. CoBots: Robust Symbiotic Autonomous Mobile Service Robots. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 4423–4429, 2015.
- [165] M. Veloso, J. Biswas, B. Coltin, S. Rosenthal, S. Brandao, T. Mericli, and R. Ventura. Symbiotic-autonomous service robots for user-requested tasks in a multi-floor building. In *IROS Workshop*, 2012.
- [166] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I. IEEE, 2001.
- [167] C. Von Hofsten. An action perspective on motor development. *Trends in Cognitive Sciences*, 8(6):266–272, 2004.
- [168] B. Wang, Z. Li, and N. Ding. Speech control of a teleoperated mobile humanoid robot. In *International Conference on Automation and Logistics (ICAL)*, pages 339–344. IEEE, 2011.
- [169] C. Wang, A. V. Savkin, R. Clout, and H. T. Nguyen. An intelligent robotic hospital bed for safe transportation of critical neurosurgery patients along crowded hospital corridors. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(5):744–754, 2015.
- [170] W.-J. Wang, J.-W. Chang, S.-F. Haung, and R.-J. Wang. Human posture recognition based on images captured by the Kinect sensor. *International Journal of Advanced Robotic Systems*, 13(2):54, 2016.
- [171] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

- 
- [172] L. Wood. Service robots: The next big productivity platform. <http://usblogs.pwc.com/emerging-technology/service-robots-the-next-big-productivity-platform>, [accessed May 9, 2018].
- [173] P. R. Wurman, R. D’Andrea, and M. Mountz. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI magazine*, 29(1):9–20, 2008.
- [174] H. A. Yanco and J. Drury. Classifying human-robot interaction: an updated taxonomy. In *International Conference on Systems, Man and Cybernetics (SMC)*, volume 3, pages 2841–2846. IEEE, 2004.
- [175] Y. Zeng, Y. Zhao, J. Bai, and B. Xu. Toward Robot Self-Consciousness (II): Brain-Inspired Robot Bodily Self Model for Self-Recognition. *Cognitive Computation*, pages 1–14, 2018.