

令和5年度博士論文

会話エージェントの言語音と身体性：
人間の印象への影響

Speech Sounds and Embodiment of Conversational Agents:
Effects on Human Impression

指導教員印	提出印

指導教員 水内郁夫 教授
東京農工大学
工学府 機械システム工学専攻
令和2年度入学
20833651

ANTONIO GALIZA CERDEIRA GONZALEZ

目次

第 1 章	Introduction	9
1.1	Thesis overview and Motivation	11
1.1.1	Overview	11
1.1.2	Speech Characteristics: Prosody and Phone choice	12
1.2	Embodiment level, experience with robots and human impression	14
1.3	Social Plantroid	15
1.4	Structure of the Thesis	19
第 2 章	Background	21
2.1	About Human-machine communication	23
2.1.1	Gibberish Speech	24
2.1.2	Prosody	25
2.1.3	Valence and Arousal	26
2.1.4	Speech Act	27
2.1.5	Statistical Bootstrapping	27
2.2	About Embodiment	28
2.2.1	Physical Embodiment	29
2.2.2	Novelty and familiarity	30
2.3	About Social and Agricultural Robotics	31
2.3.1	Plant photosynthesis, light and temperature	31
2.3.2	Soil Monitoring	31
2.4	End-to-end visual Navigation	34
2.4.1	Artificial Potential Field Method	34
2.4.2	Virtual Robot Approach	35
2.4.3	VGG-16 Deep Convolutional Neural Network	36
2.4.4	Pragmatics and Proxemics	36
第 3 章	Related Works	43
3.1	About Prosody in Human-machine Communication	45
3.2	About Embodiment	47
3.3	About Social and Agricultural Robotics	47

3.3.1	Agrobots research	47
3.3.2	Social Robotics	49
3.4	About Visual Robot Navigation	52
第 4 章	Obtaining data: the Talk to Kotaro Experiment	53
4.0.1	Talk to Kotaro: an web crowdsourcing experiment	56
4.1	Experiment Description	56
4.1.1	Structure of the experiment	57
4.1.2	Gibberish speech generation algorithm	58
4.1.3	Profile Information	61
4.1.4	Likert Scale Questionnaire	63
4.2	Platform Description	64
4.2.1	Server Side - Server Application	64
4.2.2	Server Side - Memory	65
4.2.3	Server Side - Web-page templates	65
第 5 章	Analysis of obtained data	67
5.1	Neural Network Architectures for emotion analysis	69
5.1.1	Emotion estimation from video	69
5.1.2	Sentiment analysis of recorded speech	71
5.1.3	Gibberish Speech Impression Prediction System architecture	72
5.2	Correlation Analysis	74
5.3	Analysis and results	74
5.3.1	Profile of Participants Breakdown	75
5.3.2	Impression Estimation from Video and Prosody Correlation	76
5.3.3	Analysis of the recorded speech supports the findings of the video analysis	86
5.3.4	Phone Embedding Analysis	86
5.3.5	Likert Scale Questionnaire Analysis	90
5.4	Evaluation of the GSIP	96
5.5	Discussion	97
5.5.1	Effects of Kotaro’s Gibberish Speech on Listeners	97

5.5.2	Effects of Prosody, Duration of Interaction, and Phone Choice	99
5.5.3	Performance of the GSIP System	101
第 6 章	GSIP Experiment: investigating the effects of Speech and Appearance	103
6.1	Introduction	105
6.1.1	Research questions/hypotheses	106
6.2	Experiment Plan	107
6.2.1	General Overview	107
6.2.2	Experiment Setup	110
6.2.3	Phase 1 (P_1)	111
6.2.4	Phase 2 (P_2)	113
6.2.5	Phase 3 (P_3)	114
6.3	Collected Data	115
6.4	Questionnaires	117
6.4.1	Adapted Godspeed Questionnaire for Prosody Selection Systems	117
6.4.2	Godspeed Questionnaire for Agents	118
6.4.3	Communication Systems Ranking	118
6.4.4	Agents Ranking	119
6.5	Selected Semantic-Free Utterances, and Questions and their respective responses .	120
6.5.1	Previously generated Gibberish Speech	121
6.5.2	Previously generated questions and their answers	121
第 7 章	GSIP Experiment - Execution and Results	123
7.1	Introduction	125
7.2	Experiment setup Implementation	125
7.2.1	Experiment limitations	126
7.2.2	Adapted Godspeed Scale Questionnaire	128
7.2.3	Ranking questionnaire	129
7.2.4	Emotion Analysis	131
7.3	Profile of participants	132
7.4	GSIP experiment Phase 1 - Gibberish Speech	132

7.5	GSIP experiment Phase 2 - Semantic Speech	134
7.6	Godspeed scale questionnaires response analysis	135
7.7	Responses of ranking questionnaires	138
7.8	Impression Estimation from Video	140
7.9	Results discussion	141
7.10	GSIP Experiment Phase 3 - a Study on the Effects of Embodiment Level and Novelty Bias on human Impression of ECAs.	144
7.10.1	Godspeed scale questionnaire responses	145
7.10.2	Agent ranking responses	146
7.10.3	Emotion Analysis from Video	146
7.10.4	Discussion	148
第 8 章	Development of the new Social Plantroid	159
8.1	Introduction	161
8.2	A Novel Plantroid	162
8.2.1	Hardware	163
8.2.2	Software	170
8.2.3	Detecting sunlight and shadows	175
8.2.4	Vision-based navigation system	178
8.2.5	Experiments and Results	184
第 9 章	Conclusion	193
9.1	Conclusions from “Talk to Kotaro”	195
9.2	Conclusions from the GSIP experiment	198
9.3	Conclusions from the development of Plantroid	201
9.4	Future Research	202
	Acknowledgments	204
	References	207

第1章

Introduction

1.1 Thesis overview and Motivation

1.1.1 Overview

As human societies move toward the Society 5.0 [1] paradigm and industries move toward the Industry 4.0 [2] model, smart devices are becoming increasingly ubiquitous; it is expected that more than 100 billion such devices will exist by 2050 [3]. Moreover, with the advent of virtual reality and augmented reality, there are many new possible ways of creating embodied conversational agents (ECA) – entities capable of understanding and of replying with human natural language that have a physical representation [4]. With more computing power and smaller (or no) screens, the user interface design paradigm is shifting from graphical user interfaces (GUI) to conversational user interfaces (CUI) [5]. This can already be seen in personal assistants for computers, smartphones, and smart speakers, such as Cortana, Alexa, Siri, Bixby, Google Assistant *etc.* However, such systems have one common problem: their speech sounds human-like, but they still sound unnatural because there is little to no variation on the prosody of their synthesized speech.

It should be noted that not all devices need semantic speech to convey desired messages [6, 7], such as success, failure, attention, or danger. To do this, several classes of auditory means have been used, such as sounds, music, gibberish speech *etc.*, can be employed. In particular, the use of gibberish in affective computing offers a key advantage in that it allows the communication of emotions without the need for an understandable language. This approach is useful for evaluating the effectiveness of affective prosody generation strategies as well as for implementing functional systems in a myriad of settings, since it can work across different cultures because its expressions do not depend on actual meaning of words.

Thus, Gibberish Speech is useful for evaluating the effectiveness of affective prosodic strategies as well as for implementing functional systems. This thesis investigates the effects of Gibberish Speech in a conversational setting, where humans can openly talk to robots and other embodied conversational agents without the fear of judgement, but still receiving responses that shows the ECAs are listening to what the human says and are engaged in the conversation, without actually saying anything. Gibberish speech holds a presence within popular culture, notably within movies and TV-series like Star Wars (featuring Rodian, Ewokese, Jawaese, Huttese, and other alien lan-

guages), Star Trek (where Vulcan and Klingon "languages" originated as gibberish speech in the original series) and Pingu (in order to teach children how to interpret media just from audio-visual cues). This inclusion serves to introduce an otherworldly or fantastical dimension to the narrative, effectively avoiding the need for the development of a fully coherent language. Despite its apparent randomness, gibberish speech manages to convey emotions and sentiments through character dialogues through its acoustic prosodic properties. In interactive media, gibberish speech has found its place in video games (such as Papers Please and Star Fox Command) and toys (like the Furby). These platforms utilize gibberish speech to craft immersive interactions that don't rely on conveying coherent meaning. Instead, they contribute to a unique and unconventional ambiance, enhancing the overall experience of the interaction. Despite such cultural presence, little human-robot and human-computer interaction research has been performed with gibberish speech in its center.

This way, the present thesis is dedicated to gain a better understanding on how the speech characteristics (content, phone choice and acoustic prosody parameters) and the appearance (levels of anthropomorphism and embodiment) of ECAs affect human immediate and *post-hoc* impression of the agents and the interaction with them. It also aims to develop a system capable of selecting adequate acoustic prosody parameters for the synthesized speech of ECAs, to elicit a desired emotional change on listeners. Such knowledge is of utmost importance due to the lack of general guidelines on how to develop ECAs; and, thus, aims to obtain useful knowledge that can be used to guide the development of the aforementioned entities, which are bound to become more and more common. It also showcases the development process of the Social Plantroid Robot, which was used for the necessary the investigation on speech and appearance of ECAs, and how the conclusions of the experiments shaped the development process itself.

1.1.2 Speech Characteristics: Prosody and Phone choice

When automatically generating appropriate prosody for synthetic semantic speech, it is possible to learn adequate prosody for a given speech from prosodic speech databases extracted from natural speech [8, 9] or to cause a certain impression on the listener [10, 11]. However, a similar approach for gibberish speech is not as feasible since, to the best of the authors' knowledge, there is only one emotional speech database for gibberish speech, the EMOGIB dataset [12], from which

it would be possible to extract prosodic parameters. However, since it consists only of words composed by phones present in English and Dutch, it limits the ability to develop prosody attribution systems from it, restricting the possibility cross-cultural usage. Furthermore, the emotional label present in the dataset is the perceived feeling that the utterances convey, not how they made the listeners feel.

To fill such gaps, the web-based crowdsourcing experiment “Talk to Kotaro” [13] was conducted to generate an IPA-based gibberish speech with different prosody patterns dataset labeled with the emotion change on listeners. To achieve such a goal, volunteers talked to a screen-based ECA [4] inspired by the Kotaro robot [14], which responded with gibberish speech. The developed web-site recorded both the audio of the volunteers’ speech and the video of them listening to ECA’s utterances. 33 volunteers from 8 different countries participated, speaking a total of 11 different languages; they contributed over 700 video samples. After the conversation with Kotaro, the volunteers were asked to fill out a Likert scale questionnaire, which was optional. The questionnaire was filled out by 22 participants.

Crowdsourcing data from all over the world was essential in that context, because we intend on analyzing how people from different cultures react to SFU. The initial hypothesis is that even if there are different impressions, there may be a common baseline, in similar fashion to the Bouba-Kiki effect [15]. The objective of the experiment is to gather data that will allow us to: (i) test the hypothesis that there is some common baseline on how people from different cultures react to different phones and prosody patterns and (ii) if there is a baseline, to develop a human impression prediction module using the crowdsourced data.

However, we could not find an appropriate platform which allow volunteers to talk with a robot in their web browsers while audio and video from such interactions are securely streamed to Mizuchi lab servers. The closest tools we were able to find didn’t have all the necessary features [16], were too game-like [17], introducing many other factors that could impact the impression of volunteers, or required a VR Headsets [18], which make impossible to record user facial expressions and not web-based [19]. This way, we decided to develop our own solution and make it open-source, so it can help other HRI researchers hold their own experiments online, saving time and implementation costs. It was designed in such a way that others can easily use it and modify it according to their needs. Moreover, this tool is helpful not only during times of crisis. Given that

crowdsourcing information from all over the world will make obtained datasets more diverse and, thus, research will be more robust.

The results of the analysis of the data obtained through the Talk to Kotaro experiment allowed to gain a better understanding of the effect of prosody and phone choice on human impression – the immediate emotional response – and to develop a novel bidirectional Gated Recurrent Unit (GRU) neural network architecture that is capable of estimating human impression for given phone and prosody inputs, called Gibberish Speech Impression Predictor (GSIP). The performance of the system was validated by an in-person experiment with 28 participants; and achieved good precision on estimating the impression caused by the gibberish utterances of three distinct ECA.

1.2 Embodiment level, experience with robots and human impression

Embodied Conversational agents can thus have different levels of physical embodiment [20]; some agents are just text on a display or a voice that speaks to users, while others are robots that are fully capable of sensing and interacting with the world around them. Since their main function always has a social component, in the sense that they perform tasks where it is necessary to talk to humans, it is important to understand how the level of physical embodiment relates to human perception. If this is properly understood, it is possible to design an agent with a level of embodiment that is sufficient, since higher levels of embodiment tend to make an ECA more expensive, i.e., a speaker connected to a computer is cheaper than a robot with a plethora of sensors and actuators necessary to perform its tasks. Moreover, since conversational agents can occupy multiple displays or robot bodies at once, change bodies as needed, and share bodies with other ECAs, it is important to know how changing the level of embodiment changes human opinion about its capabilities.

Studies have already been conducted to investigate the relationship between physical embodiment level and human engagement, human perception of ECA [21, 22], and how well users perform certain tasks when interacting with agents of different embodiment levels [23, 24, 25]. However, there is a possibility that such results are due to novelty preference - the preference for new experiences - which may have played an important role in these results. Previous works,

even when acknowledging this possibility, have not investigated the effect of novelty preference on participants' impressions and preferences.

To address this gap in the literature, an experiment in which participants have conversations with three different versions of the Social Plantroid robot was performed: a 2D avatar displayed on a computer screen (referred to as a screen agent), a 3D model displayed on a Gen 1 Looking Glass 8.9-inch holographic display (referred to as a holographic agent), and the real Plantroid robot (robot agent). All ECAs use the same GPT-3-based chatbot and eSpeak speech synthesizer – their appearance, shown in Figure 7.2, is the only distinguishing factor. The executed experiment was the first human-robot and human-computer interaction study to use a holographic display for 3D visualization of embodied agents. After interacting with an agent, volunteers were asked to complete a short adapted Godspeed Scale questionnaire and, after interacting with all agents, to rank which agent they preferred to interact with. All volunteers had to fill out a profile with relevant personal information and their experience with robots, which was used to assess how novel the experience of interacting with the Plantroid robot was.

Current understanding was that, although not causing significantly better performance of users, higher levels of physical embodiment of ECA seems to cause a higher engagement [22, 23]. The proposed experiment was conducted to challenge this understanding and thus verify the following research hypothesis:

H_1 : Novelty preference plays a strong role in engagement in interactions with ECAs.

If hypothesis H_1 is true, we expect a negative correlation between participants' level of experience with robots and their impressions of the robot agent.

Thus, three main contributions of the experiment were: 1) investigate the correlation between volunteers' impressions and the level of physical embodiment of ECA through questionnaires and emotion estimation from video; 2) investigate the correlation between novelty preference in the preference of ECA and 3) it is the first study to use holographic displays in an HRI experiment.

1.3 Social Plantroid

In order to test, validate and demonstrate the novel embodied conversational agents paradigms created through the present research and the developed systems, a novel version of the Plantroid(Plant+droid)

family of robots [26, 27] was developed, called Social Plantroid, since it joins the smart agricultural worker aspect of previous Plantroids with a social side, which aims to transform plants into pets and friends.

In the Agriculture 4.0 paradigm, the latest technological advances in fields such as internet of things , big data, machine learning, remote sensing and precision farming are put together in order to optimize crop yield and quality, while minimizing the environmental impact, costs and intensiveness of labor [28]. Moreover, given the current trend of urbanization [29, 30] and that many countries have an ageing population [31], the reduction of the available workforce and of the average farm was only a natural, albeit perilous, outcome [28, 29, 32]. In that sense, robotics is expected to play a central role in future of farming due to its resource saving, precision improving and labor saving potential [28]. Interest on agrobots (agricultural robot) research has, thus, only grown in the last few decades [33].

With the reduction of available farmland, greenhouse farming [29, 33] and urban agriculture [34, 32, 35] appear as very labor intensive solutions, which require precise resource management. Research on IoT, big data machine learning, AI-assisted decision-making systems address the resource management aspect [36]. The original Plantroid research [26, 37], and by extension this present work, are inserted in the corpus of robotics-based labor-saving solutions research. However, whereas previously developed Plantroids , hereby referred to as Plantroid Omni [26] (shown in Figure 1.1a and Plantroid mini [37] (shown in Figure 1.1b only address the labor intensive problem of carrying plants into and out of sunlight in smart-greenhouses and plant factories; the novel Social Plantroid (shown in Figure 1.1c also takes care of monitoring the soil of the plant, information management and communication.

However, much more relevant to the scope of this thesis, Social Plantroid is an embodied conversational agent that aims to help you take care of plants and wants to be a companion, notifying the necessities of the plant when its sensors detect that the soil lacks nutrients, or needs watering and holding complex and engaging conversations through a novel GPT-J-based Dialogue Management system that takes Pragmatics and Proxemics principles into account for communication. It can also uses the developed Gibberish speech and prosody generation techniques, coupled with a powerful emotion and facial expression framework to express the needs of the Plant it carries. The development of the Social Plantroid robotacknowledges the fact that, while Robots and AI might

substitute human labor in certain conditions [38, 39], taking care of plants has psychological benefits to humans and, thus, Plantroid should only help instead of substituting the care-taking labor. This way, as it is expected for many robots [40], Plantroid works together with humans. Robots in a cooperative setting need to competently communicate, and to understand human verbal and non-verbal communication [41, 42]. Its pet-like appearance was also chosen to make the robot appealing [43] for home-owners and further contribute to the impression that the plant became a pet.

Previous Plantroid versions required an external camera for environment navigation and, most importantly, for performing its main task of finding sunlight or shadow, accordingly to the need of the plants they carried. The novel Social Plantroid has two cameras, one gray scale OMRON B5T-007001-010 [44], used for human detection, emotion recognition and sunlight detection and an Adafruit MLX90640 IR Thermal Camera [45] for sunlight detection.

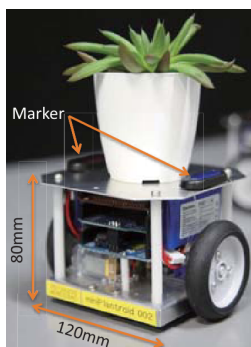
This way, the novel Social Plantroid was developed to be a Human-Robot Interaction research platform and an agrobot research platform. It addresses the problem that doomed many social robots to fail as products: the lack of perceived utility by customers [46]. It is, to the knowledge of the authors, the first open-source agricultural robot with a social function. Other novelties presented in this research is that a simple, but effective, sunlight-seeking algorithm which requires no external cameras was developed, together with a VGG-16-based end-to-end visual navigation architecture that allows Plantroid to avoid obstacles while seeking the best sunlight or shadow.

Vision is a powerful sense for navigation – it can be used to achieve the 4 principal tasks of robot navigation: localization, mapping, path planning, and locomotion [47]. Moreover, since most humans rely on vision to navigate in their daily lives, interest in vision-based robot navigation research is natural. This work is inserted in that context, focusing on end-to-end locomotion and obstacle avoidance using monocular gray scale images to estimate the heading direction of a differential drive Plantroid robot.

Initially, most approaches for visual navigation were based on Image Processing [47] techniques, but as computing power increased and machine learning matured, machine-learning-based solutions became more present in the field. However, one gap in end-to-end neural-network vision-based navigation research is that no solution was developed to directly learn the behavior derived from using the Artificial Potential Field (APF) [48] method for path planning, the



(a) Autonomous Movable Fruit Growing Plantroid.



(b) Plantroid Mini.



(c) 3D render of the Novel Social Plantroid .

Fig. 1.1: The Plantroid Family.

same method used in the previous Plantroid models for seeking sunlight, while avoiding walls and other robots [27].

Vision-based end-to-end methods have the advantage of eliminating the need for robot localization and mapping the environment, generating locomotion decisions by directly sensing the environment, a behavior known as reflex approach [49]. That allows to reduce necessary computation power and reduce the number of necessary sensors, making robots cheaper, lighter, smaller and more energy efficient. Thus, in this work, localization and mapping problems are not addressed; and it assumes that for the task of seeking sunlit areas, Plantroid encoders are precise enough, since the objective destination is an area far larger than the robot itself. Knowing the map is not essential, since the architecture successfully learns how to avoid walls, static and mobile obstacles.

Works [50, 51, 52] have used monocular images to estimate the distance of obstacles from the robot and then used variations of the APF method for trajectory planning. The proposed VGG-16[53] based architecture yields the robot heading directly from images, eliminating the need of running the APF method while achieving comparable performance. APF was chosen as the planning method for training data generation for the proposed architecture because Social Plantroid's main navigation goal is to move into and out of sunlight while avoiding obstacles. The

intensity of the sunlight also translates well into the attractive potential of the robot's goal, as it was done for the previous model, albeit from an external camera. Moreover, its implementation is simple and has many variants. Trajectories planned through APF method are followed through the virtual robot approach, which is also easily implemented.

The proposed architecture consists of a convolutional neural network (CNN_1), which receives RGB images from the camera as input; and a multilayer perceptron (MLP_1), which receives the current location (x_c, y_c) of the robot and the desired robot destination (x_f, y_f) . The outputs of the aforementioned neural networks are concatenated and used as the input of a second multilayer perceptron (MLP_2), which then outputs a predict way-point (x_n, y_n) for a time horizon T , which should enable the robot to avoid all moving and static obstacles.

Using the APF method and the virtual robot approach, a little over 30h of simulations were run in 3 distinct environments: a house, a cafe, and a meeting room, where the robot navigates from an initial position $p_i = (x_i, y_i)$ to a final position $p_f = (x_f, y_f)$ while avoiding mobile and static obstacles. Every second, an image is saved, together with the current robot pose $(x_r, y_r$ and $\theta_r)$, current destination, and the future heading of the robot obtained through the aforementioned techniques. Generating data, training, and evaluating the navigation architecture in a simulated environment allows for cheaper and faster development since it does not wear out real robots, does not require modifications to the environment and does not need to run in real-time.

1.4 Structure of the Thesis

The structure of the present thesis is as follows. Chapter 1 introduces the main objectives, gives basic context and rationale for the research that was performed and outlines the structure of the thesis. Chapter 2 introduces the many necessary concepts for understanding this thesis, while Chapter 3 introduces previous research that is related to the investigations performed in this thesis, while outlining the difference between the current and previous works. Chapter 4 introduces the web crowdsourcing platform that was developed, together with the experiment that was performed to gain a deeper understanding on how phone and prosody choice affects human impression, also introducing the Gibberish speech generation algorithm and the Likert Scale questionnaire used to obtain a deeper understanding of why participants reacted the way they did in the experiment. Chapter 5 introduces what methods were used for analyzing the data obtained through the Talk to

Kotaro experiment, showcases and discusses its results and introduces the Gibberish Speech Impression Prediction system. Chapter 6 introduces the GSIP experiment, which aimed to investigate the performance of the developed system and elucidate how volunteers perceived gibberish speech, English, distinct prosody selection methods and hope the levels of embodiment and anthropomorphism of Embodied conversational agents impacted their impression. Chapter 8 introduces the development process of the new Social Plantroid robot and its many systems developed for human robot interaction. Moreover, it also introduces its end-to-end VGG16-based architecture for visual navigation and obstacle avoidance in social environments. Finally, Chapter 9 outlines the conclusions from the present thesis and future works.

第2章

Background

This chapter provides essential information to understand this thesis. It is divided into 4 different domains of knowledge that were necessary to perform all research topics present in this thesis – human-robot and human-computer interaction are very interdisciplinary research fields. First, Section 2.1 presents knowledge related to phonetics, speech synthesis and multimodal estimation. After that, Section 2.2 introduces Embodiment levels and Novelty bias, or preference for novelty. Section 2.3 presents knowledge necessary for visual navigation of robots, pragmatics and proxemics, soil monitoring and plant health estimation.

2.1 About Human-machine communication

This section briefly introduces concepts necessary for understanding this work and presents related work. Its subsection 2.1.1 explains in detail what Gibberish Speech is and introduces the International Phonetic Alphabet, whose symbols serve as building blocks for Gibberish Speech in our work. Subsection 2.1.2 explains what prosody is in the context of linguistics. Subsection 2.1.3 explains the valence-arousal emotion classification model. Subsection 2.1.4 explains the mathematical model that maps the listener’s emotional response to an IPA phone, prosody pair.

In the field of HRI, the topic of Human-robot communication is of great interest, because seamless cooperation between humans and machines are essential in the Industry 4.0 [2] and Society 5.0 [1] paradigms. Verbal communication is one of most natural means of information exchange between humans and thus, this is one of the focal points of the research area. However, communication does not always need to be done with intelligible words. Emotions can be conveyed by SFU [54], which is reflected in many pop-culture icons like R2-D2, Wall-E and EVE, which do not need meaningful utterances to convey their message and capture hearts. SFU are traditionally classified in 4 types: Gibberish Speech (GS), Non-Linguistic Utterances (NLUs), Musical Utterances (MU) and Paralinguistic Utterances (PU) [55]. This work focuses on GS, which are defined as meaningless utterances that are composed from real phones, resembling human languages. For that similarity, many HRI studies have been performed, trying to better understand how such SFU impacts human impression, the experience of interacting with the robot, if the message is being correctly understood; among many others, such as [56, 12, 55].

HRI, as a research field, has the problem of measuring human thoughts and feelings regarding an interaction with robots as a way of validating research hypothesis and developed technologies

on its core. Unfortunately, there is still no sensor capable of that. The most used methods of trying to gauge the internal state of test subjects are questionnaires, due to their ease of implementation. Works [57, 58, 59], for example, have used questionnaires to measure the overall experience of test subjects. However, such methods have several limitations, such as research subjects trusting scientists too much (or too little), the fact that the questionnaire is not immediately answered as the interaction happens and that the precision of any conclusions drawn from responses depends on the correct statistical knowledge of the researchers [60].

To avoid these limitations, a vision-based human impression estimation [61] system was employed. There are several different emotion estimation techniques, such as Bayesian Networks, SVMs, Decision trees [62], Deep reinforcement Learning [63], Deep CNN [64] among others. In the scope of this thesis, three techniques are used. VGG-16 and Resnet-18 Deep Neural Networks are used to estimate human emotion in terms of valence and arousal for the present facial expression. Moreover an OMRON HVC-P2 B5T-007001-010 camera is used to obtain an emotion label of the human, allowing Social Plantroid to change its behavior, facial expression and prosody without needing to run the heavy neural network models.

Regarding the Gibberish Speech generated in this work, IPA symbols [65] were used to build it, allowing the Embodied conversational agents to speak phones from every language IPA is capable of representing. This is similar to Hanamogera [56], which uses Japanese Language phonemes to generate gibberish speech. No other Gibberish Speech HRI Research has been performing using the IPA for its utterances, which is a novelty provided by this research.

All embodied conversational agents in this thesis used *espeak* [66], a formant synthesis speech synthesizer, to have a voice. *Espeak* was chosen due to its easy of use, small size, quick synthesis of clear speech and, most importantly, the fact that it accepts IPA input (in the ASCII-IPA form) and it allows us to control the prosody parameters of the generated speech: volume, speed and pitch.

2.1.1 Gibberish Speech

In human-computer communication, when a given language is used for communication, it limits the set of people who can effectively understand what an ECA is trying to convey; and the meaning of words can have multiple interpretations that affect the impact on a listener. To avoid

such limitations, semantically free utterances have been used to convey emotions such as anger, sadness *etc.* Such utterances use musical cues such as tempo and pitch to convey emotion. For example, sounds with slower tempo, lower pitch, and little variation convey sadness, while sounds with faster tempo, high volume, and intensity can convey anger. Such cues also apply to human-like speech, allowing it to convey such emotions without conveying meaning, but still resembling a language. There are four main classifications of semantic-free speech: [6]: (i) gibberish speech, SFU, which are composed of human speech sounds; (ii) paralinguistic utterances, which are composed of human non-speech sounds, such as laughs, sighs *etc.*; (iii) musical utterances, which use musical sounds to convey messages and feelings; and (iv) non-linguistic utterances, which consist of beeps, whirs, and pings, among many other sounds to communicate [6].

This work focuses on gibberish speech (GS) because it is useful for systems that do not require meaningful vocalizations to convey certain meanings, such as in human-robot interaction, video games, or animation. It can also be beneficial when users need to communicate with a technology that has little natural language processing capability, such as voice-activated devices with low processing power. The ability to convey more subtle emotions and intentions through fluctuations in pitch, rhythm, and other acoustic aspects is an advantage of using gibberish speech over other SFUs.

2.1.2 Prosody

In linguistics, prosody is defined as the study of larger units of speech, such as syllable characteristics, intonation, stress, and rhythm [67]. Listeners can infer the emotional state of speakers from the prosody of their utterances, since someone who is excited, for example, may speak faster, louder, and at a higher pitch than usual.

The most important auditory variables in prosody are pitch (how low or how high the voice is), rate (the length of the utterances), loudness (how loud the voice is), and timbre (the quality of the sound of the voice) [67]. This work is concerned with the first three characteristics, assuming that decreasing the quality of the audio will lead to negative reactions because it will make it harder to understand what the ECA is saying.

2.1.3 Valence and Arousal

The question of how many human emotions there are and how to classify them is an important problem in psychology, and thus, many classification models have been developed. One such model is Russell’s two-dimensional model of valence and arousal [68], which classifies emotions in a continuous valence–arousal space. Valence represents how positive or negative an emotion is, while arousal represents how aroused a person is from relaxation to excitement [68]. The valence–arousal emotion space is defined over $\{v \in \mathbb{R} \mid -1 \leq v \leq 1\}$ and $\{a \in \mathbb{R} \mid -1 \leq a \leq 1\}$, which produces the emotion space shown in Figure 2.1, along with the positioning of some emotions. This model is often used because it produces a continuous emotion space rather than discrete labels, such as Paul Ekman’s six or seven basic emotions [69] or Plutchik’s wheel of emotions [70]. It is often used for emotion estimation from facial expressions, the same context of this work [71].

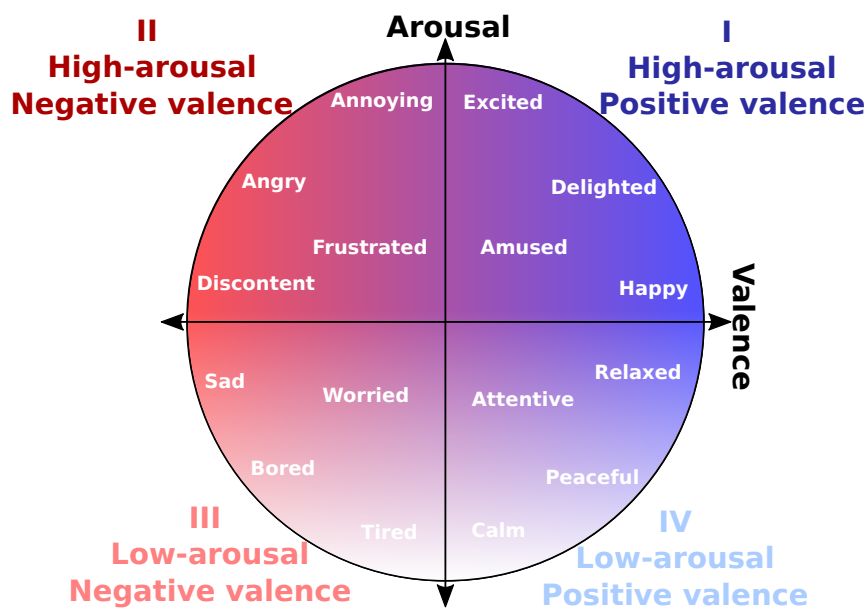


Fig. 2.1: Russell’s two-dimensional model of valence and arousal and the mapping of some emotions in it.

The emotional state of a person at a given time t is then defined as $E_t = (v_t, a_t)$.

2.1.4 Speech Act

Speech is an act on itself with effects on listeners. Since the speech used in this work is meaningless, the only effect it can have on listeners is emotional, and thus this work only considers perlocutionary acts and studies their perlocutionary effects on human listeners. A speech is defined as $S(w, P)$, where w is a vector containing each phone to be spoken and P is a $\|w\| \times 3$ matrix containing the prosody (volume, speed, and pitch) associated with each phone. An example of a speech is $S_{example}$:

$$S_{example} = S([\text{b}, \text{a}, \text{t}], \begin{bmatrix} 100 & 130 & 45 \\ 90 & 130 & 50 \\ 95 & 140 & 50 \end{bmatrix})$$

The communication act by the ECA can be defined as $C[S(w, p)] = f(S(w, p))$, where f is a rendering function, which, in this work, represents the eSpeak speech synthesizer. Even if listening to an utterance does not lead to an action, it is expected to produce an impression \vec{I}_S , which is defined as $\vec{I} = \vec{\delta}_E(\delta_v, \delta_a)$, representing the change in valence and arousal caused by the speech act S . This change can be modeled as $E_{t+1} = g\{E_t, C[S(w, p)]\}$, where t is the moment before hearing S and $t + 1$ is the moment after; and g is a function representing how a listener responds to utterances. This function represents individual preferences, sensibilities, cultural background *etc*; and is very difficult, if not impossible, to model. However, with enough data, it is possible to learn listener preferences for phonetic and prosodic choices through machine learning.

2.1.5 Statistical Bootstrapping

Bootstrapping is a statistical technique introduced in the late 1970's that enables researchers to make data-based inferences without strict distributional assumptions for univariate and multivariate data. It involves two distributions: the underlying distribution of the data (for example, normal or binomial) and the distribution of a computed statistic (in our case, Stuart–Kendall τ_C correlation). New m data sets are formed by Monte Carlo resampling, each one containing the same number of observations n as the original data set [72]. Monte Carlo resampling is performed through randomly selecting points in the original data set and copying them into the new data set, until there are n points in the new data set.

That way, a data sample of the original data set might appear one time, multiple times, or not at all in the new data set. Such an operation is performed m times; and for each new data set created, it is necessary to perform the computed statistic operation. Now, we have a distribution of computed statistic results, from which we can obtain a confidence interval through several techniques, such as percentile [73], bias-corrected (BC) [74], bias-corrected and accelerated (BCa) [75], and approximate bootstrap confidence (ABC) [76], among others.

The authors have chosen to use the percentile method since it suffices for the performed analysis and due to its easiness of implementation. The percentile method consists of plotting the frequency histogram of the m computed statistics of the new sampled data sets, and the 95% confidence interval will consist of the values between the 2.5th and 97.5th percentiles. These percentiles represent the lower and upper bounds of the confidence interval, respectively. The resulting confidence interval yields a range of values within which the true parameter value is likely to fall with a certain level of confidence, in our case, 95%.

The idea behind such process is to estimate the sampling distribution of a statistic by repeatedly sampling with replacement from the observed data. The ultimate goal is to make inferences about a population parameter or the distribution of a statistic even when you have a limited amount of data, as long as the distribution of the limited data set somewhat resembles the real distribution of the real-world variable.

2.2 About Embodiment

Embodiment is an idea of great interest in the fields of psychology, cognitive science, and human-robot and human-computer interaction research. The idea of embodiment is based on the concept that cognitive and perceptual processes cannot be separated from the physical body through which the mind interacts with its environment. Thoughts, emotions and perceptions, according to this conception, are shaped by bodily processes and experiences, i.e. the body is not a simple passive container for the mind, but plays an active role in shaping its mental processes [77]. However, the exact definition of embodiment, to what extent the mental processes are embodied, is not exactly agreed, since there are a few different definitions and models [20, 78, 77].

However, since this work is concerned with the effect of the level of how physical the body of an ECA is, the only concept of interest regarding embodiment is physical embodiment as defined

in [20]. The screen agent has a colorful 2D body, while the holographic ECA has all 3 dimensions, even though it has no physical body that can interact with the environment and be touched by users. In this sense, the robot has the highest level of physical embodiment because it is a physical entity. In another sense, all agents have very similar capabilities, since they all use the same microphone and the same camera to record the volunteers, and use the same voice to speak, and use the same GPT-3-based chatbot [79].

2.2.1 Physical Embodiment

Since the definitions of embodiment and physical embodiment are still not so clear, it is important to establish a framework that allows useful conclusions to be drawn from the data obtained in our experiment. We propose a 4-level scale of physical embodiment, which is illustrated in Figure 2.2:

- 0) No representation
1. 2D representation
2. 3D representation
3. Physical body

The proposed 4-level scale of physical embodiment is by no means definitive - it is a tentative framework to make this work more understandable, since the boundaries between the more "real-like" agents are very clear. However, the boundaries between levels 1 and 2 may not be so clear. For example, if we compare a photorealistic 2D representation of the actual robot body with an abstract 3D representation of the robot, which should be considered more or less embodied? There is no obvious answer to this question, but it is very clear that the actual robot body is the most embodied agent. Scientific research needs to be done to try to establish clearer boundaries between levels 1 and 2, but apart from such problems, the proposed framework is useful for understanding the present work.

The researchers are aware that the communication medium used by an ECA also contributes to the impression of its physical embodiment. That is, an agent that displays text may feel less

physically embodied than one that uses speech; and an agent that uses gestures to emphasize its speech may feel more embodied than one that only speaks. However, since these differences say more about the capabilities of the system than about how close the embodiment is to a fully real and physical body, the researchers feel that it would be appropriate, at least in the context of this thesis, to consider such capabilities as a separate dimension of embodiment. This works for the present research because all agents used the same communication medium - speech synthesized by the eSpeak [66] speech synthesizer.

2.2.2 Novelty and familiarity

Novelty preference is a well-known psychological effect that is thought to stem from the biological need to understand new events in our lives in order to increase our chances of survival. This instinct leads people to seek to understand and be fascinated by new experiences. However, the opposite effect is also observed; humans also prefer familiar events because they are already understood and known to pose no threat [80]. In the context of human-robot interaction, previous studies have found that users tend to show higher engagement for ECAs that are fully embodied; and we suspect that this phenomenon is related to novelty preference, since most people do not have extensive experience interacting with robots. This effect also extends to many other occasions, not just HRI.

Four levels of familiarity with robots were established by analyzing the level of experience of many volunteers who participated in the experiments executed to validate the performance of the developed prosody selection techniques and to investigate the effects of embodiment and experience with robots had on the impression of volunteers about conversational agents:

- 0) No experience: no previous interaction with robots;
1. Beginner: few brief interaction with robots without communication;
2. Intermediate: multiple interaction with robots or a few with communication;
3. Experienced: specialist in robots or someone who has had extensive interaction with robots.

None of the participants had previously interacted with holographic displays; and all of them had previously interacted with monitors, even if they had not previously interacted with screen-

based conversational agents. Thus, the novelty factor of the experiment is largely controlled by prior experience with robots, which is used to analyze novelty preference in the experiment.

2.3 About Social and Agricultural Robotics

This Section provides background information necessary for understanding some of the design choices and algorithms employed in Social Plantroid . Since the main agricultural function Plantroid performs is guaranteeing that the plant it carries receives enough sunlight, Subsection 2.3.1 describes the photosynthesis mathematical models which are used to decide whether the robot should seek sunlight or shadow and how intense the sunlight needs to be. Moreover, since Social Plantroid monitors the pH, Moisture. Finally, regarding the Social Robot side, the design principles which guided the development of the conversation engine are explained in 2.4.4.

2.3.1 Plant photosynthesis, light and temperature

Plant photosynthesis is the organic process through which plants harvest energy from sunlight, converting it into biochemical energy, which is used to sustain its many physiologic processes ([81]).

Light is one of the principal driving factors in photosynthesis; without enough sunlight, a plant will not be able to develop properly, accumulating less dry matter than otherwise. However, excessive sunlight is also harmful, leading to photoinhibition, photooxidation and damage to the leaves, causing an early maturation of the vegetable and reduced dry matter accumulation ([82]).

Moreover, when the temperature of the environment is high, some enzymes involved in the mechanism of Chl biosynthesis decrease, *e.g.* the 5-aminolevulinatase (ALAD)[83]. This way, Plantroid also needs to seek shadow whenever the measure temperature exceeds a certain threshold.

2.3.2 Soil Monitoring

The overwhelming majority of plants acquire the majority of their nutrients through their roots from the soil; with carnivore plants and air plant being notable exceptions. Thus, in order for a plant to enjoy a healthy growth and life, the soil must be fertile, that is, contain the necessary

nutrients for an specific plant species. The Social Plantroid , besides moving the plant into and out of sunlight, monitors many import contents of the soil, so the owners of the plant may take corrective action in order to protect the plant.

Important soil characteristics are moisture, pH, salinity and NPK contents, and the present Sub-section is dedicated on outlining the importance of each one of them for the healthy development of plants.

Moisture

Plants, like every living being, require water for many of their physiologic processes, including photosynthesis. Most plants absorb necessary water from the soil by osmosis through their roots and, thus, guaranteeing that the soil has enough water in it is essential. However, if there is an excess of moisture, it may deprive the roots of oxygen, killing the plant. Measuring the moisture levels in the soil and ensuring that it stays in a healthy range is, thus, essential for obtaining optimal plant growth.

Soil pH and photosynthesis

The relationship between a variation in the pH of the soil and the CO_2 absorption by the plant is approximately linear ([84]):

$$\Delta_{A_{CO_2}} = -K_{pH}\Delta_{pH} \quad (2.1)$$

Where $\Delta_{A_{CO_2}}$ stands for the variation on the carbon dioxide assimilation by the plant, K_{pH} is a constant specific to each species of plant, indicating how sensible it is to changes in the pH of the soil and Δ_{pH} denotes the change of the pH of the soil.

It is necessary to note that every plant species has a maximum CO_2 absorption rate and, thus, decreasing the soil pH after a certain value will not lead to an increase the plant's CO_2 absorption, but actually the opposite, since the increased presence of H^+ ions in the soil will damage the roots of the plant, reducing the absorption of water and mineral nutrients (such as Nitrogen, Phosphorus and Potassium, whose concentration in the soil are measured by Plantroid). Such lack of water and nutrients will eventually lead to a complete inhibition of photosynthesis in the plant ([85]).

Salinity

Soil salinity is defined as the quantity of salts dissolved in the aqueous phase of the soil. The quantity of such salts may negatively impact the plants' health, since a high concentration of salts in the soil might cause soil acidification. That will, in turn, stunt the growth of the plant up to over 90% ([86]). Moreover, if the salinity of the soil is too high, the osmotic process through which plants uptake water from the soil will reduce and, possibly, revert, that is, the roots might lose water to the soil.

Thus, it becomes very important to monitor the salinity of the soil in the pot, since the water used to water the plant will very probably contains dissolved salts in it. One of the easiest and cheapest ways of measuring the salinity of the soil is to measure its electrical conductivity (henceforth referred to as EC), since conductivity of the aqueous phase of a soil increases along with its salinity.

Nitrogen, Phosphorus and Potassium (NPK)

Nitrogen, Phosphorus and Potassium are very important nutrients for the development of plants, since they are used in many of the physiological processes of a plant. In a potted plant, the plant removes such minerals from the soil during its life and the soil will eventually become deprived of such components. The health of the plant will, then, be negatively impacted by lack of nutrients. This way, efficient fertilization is essential to keep a plant healthy and knowing which components and how much is necessary helps that process.

Plants require a minimum amount of such nutrients to be able to develop and, since plants do not grow indefinitely, there is an upper limit of the uptake of NPK from the soil. This way, the curve which better represents the concentration of a nutrient in the soil vs the expected final dry mass Y of the vegetable is expected to be a Sigmoid curve, which is obtained through the following logistic function ([87]):

$$Y = \frac{A}{1 + e^{(b-cM)}} \quad (2.2)$$

Where A is the maximum biomass a plant species can achieve in very ideal conditions, b is the intercept parameter and c is the nutrient response coefficient (how much the plant's growth

depends on such nutrient).

It must be noticed that the values of A , b and c have been obtained empirically for many species of plants for N, P and K. Moreover, it is important to highlight that after a certain amount, the presence of such nutrients might actually be harmful for the development of the plant and, thus, the Sigmoid curve does not appropriately model the effects outside the “healthy range” for the vegetable in question.

2.4 End-to-end visual Navigation

In order to achieve autonomous robot navigation, many systems are necessary; in this work, a heading direction system (either the APF method navigation planner coupled with the virtual robot approach or the proposed VGG-16-based architecture) and the low-level locomotion controller, which ensures that the robot will turn into the heading direction given by the heading direction system.

This Section, thus, briefly explains the necessary elements for generating the training data and the VGG-16-based architecture: (2.4.1) Artificial Potential Field Method, (2.4.2) Virtual Robot Approach, (2.4.3) VGG-16 Deep Convolutional Neural Network and, finally, of (3.4) Related Works.

2.4.1 Artificial Potential Field Method

The artificial potential field method [48] is an extensively studied path-planning method for mobile robots. In its original version, obstacles exert a repelling force on the robot, while a target destination exerts an attractive force, as shown in Figure 2.3. The resulting force of all obstacles and goals will move the robot into its next position, and this process will continue until the final destination is reached. The method has been used widely for its easiness of implementation, but it does not guarantee that a trajectory will be found even if it exists. Such a problem happens at points where all attractive and repulsive forces cancel out, leaving the robot stuck. Many variations of the method have been created to tackle such limitations, but they are beyond the scope of this work.

With the path planned, it is necessary to implement a trajectory following the control algorithm, which is explained in Subsection 2.4.2.

2.4.2 Virtual Robot Approach

The virtual robot approach [88] is a path-tracking algorithm, which considers that there is another robot at a reference point in the trajectory, the virtual robot, which is governed by a differential equation containing the feedback error. This gives the virtual robot its own dynamics, moving along the path in a way that the real robot is able to mimic. The approach is robust to measurement errors and external disturbances because the motion of the virtual vehicle is governed by the tracking error feedback, which makes possible to use only proportional controllers for the real robot. The idea behind the Virtual Robot Approach is shown in Figure 2.4.

The objective is, then, finding a lateral control $\delta_f(t)$ and longitudinal control $v(t)$ which makes the robot follow the reference trajectory smoothly. The trajectory is parameterized by the virtual vehicle $s(t)$, moving in the trajectory defined as $(x_d, y_d) = (p(s), q(s)); 0 < s \leq s_f$. The subscript d stands for desired and the subscript f stands for final. It is assumed that $p'^2(s) + q'^2(s) \neq 0 \forall s \in [0; s_f]$. The control objective is as follows:

$$\lim_{t \rightarrow \infty} \sup \rho(t) \leq d_\rho; \quad (2.3)$$

$$\lim_{t \rightarrow \infty} \sup |\psi - \psi_d| \leq d_\psi; \quad (2.4)$$

where $\rho(t) = \sqrt{\Delta x^2 + \Delta y^2}$, $\Delta x = x_d - x$, $\Delta y = y_d - y$, ψ is the current heading of the real robot and $\psi_d = \text{atan2}(\Delta y, \Delta x)$ is the desired heading angle of the robot. Additionally, $d_\phi > 0$ is a small number that depends on the maximum curvature of the given trajectory and d_ρ is the look-ahead distance for the virtual robot.

Thus, the robot's linear speed is given by $v = \sqrt{\dot{x}^2 + \dot{y}^2}$ and, because the robot is assumed to move slowly through the environment, it can be considered to be constant, making it necessary only to control the robot's heading. Finally, the control law for the heading of the robot is given by:

$$\delta_f = -k(\psi - \psi_d), k > 0 \quad (2.5)$$

2.4.3 VGG-16 Deep Convolutional Neural Network

Convolution Neural Networks (CNN) are standard artificial neural networks able to process multidimensional data. This way, because they are capable of processing spatial data through shared weights, CNNs excel at tasks such as image recognition and processing and computer vision.

The VGG-16 CNN used in this work is a deep CNN (DCNN), which was originally developed for large-scale image recognition, that is, a single neural network is capable of recognizing a large quantity of distinct objects [53]. Thus, it is able to recognize distinct environments and obstacles while navigating; the reason why such DCNN was chosen.

2.4.4 Pragmatics and Proxemics

The Social Plantroid has a simple, albeit effective, conversational engine which was designed with modularity in mind. At the center of that conversational engine is a chatbot, whose dialogues, albeit simple, allows Plantroid to hold enjoyable conversation with its owners, while announcing the needs of the plant whenever necessary. The design principles used for writing Social Plantroid's dialogues and conversational behavior were based in Neo-gricean Pragmatics and Proxemics, which are briefly introduced in Subsubsections 2.4.4 and 2.4.4, respectively.

Pragmatics

In linguistics and in other communication-related fields, Pragmatics is considered to form a triad with Semantics and Syntax. While syntax studies the formal relations between communication signs and semantics dedicates itself to studying how signs relate to objects and actions in the external world; pragmatics, investigates the relation between signs and those who interpret them, that is, language users. In the context of social robotics, language users might be either humans or robots and signs have a rather broad definition; not being limited to written, drawn or spoken signs, but including bodily gestures and facial expressions [89].

Since Pragmatics studies the interpretation of signs, it implies that a collection of signs has reached an individual; that is, it presupposes exchanges between language users in some way. Thus, language users cannot be seen as isolated beings, but as social creatures in the Aristotelian sense; being inserted in sociocultural contexts which might change interpretation of given signs. Moreover, through the use of verbal and nonverbal signs, language users are also able to interact with and change the aforementioned social contexts.

In the aforementioned signs exchanges, it is expected that there will be a certain extent of cooperation, of "good will", between language users. That is to say, they will avoid telling lies, say irrelevant things, avoid ambiguity, wait their turn to speak, *etc.* This assumption is called Grice's Cooperative Principle, in which a few maxims can be observed or breached. The Gricean maxims are as follows:

- Quantity

1. Contributions to the conversation are as informative as required;
2. Do not make your contributions more informative than is required.

- Quality

1. Do not say what you believe to be false;
2. Do not say that for which you lack adequate evidence.

- Relation

1. Be relevant.

- Manner

1. Avoid obscurity of expressions;
2. Avoid ambiguity;
3. Be brief;
4. Be orderly.

If a violation of a maxim is blatant and intentional, it is called a flout. If an individual is caught flouting, it drastically changes how others interpret his or her signals, because they will try to find the real intentions of the flouting person, looking for a layer of underlying meaning called implicature. That phenomenon is specially important for robots and chatbots, because it has been observed that humans will try to read more into what the machine is trying to say than it was originally intended by its creators ([90]).

Expecting humans to follow Grice's Cooperation principle might be too generous of an assumption, but since a service robot is only useful if it serves its owners, any communication from the robot should respect such principles, that is, the robot is always cooperative. Moreover, since the robot is taking care of a plant which belongs to someone, we can assume that it is in the best interest of that someone to cooperate, for the sake of the plant.

The robot also needs to be polite in its conversation. It can be noticed that "be polite" is not included in the Gricean maxims, but in many sociocultural contexts, it is essential for establishing cooperation. For example, in a context where a subordinate talks to a superior, honorifics are necessary, violating the maxim of Quantity, by adding words which do not contribute to the information that is being exchanged. To consider that necessity in exchanges, Neo-gricean pragmatics, thus, includes Theory of Politeness, adding the following politeness maxims [91]:

1. Tact: avoid damaging reputation and negative implications;
2. Generosity: selflessness;
3. Approbation: minimize criticism, maximize praise;
4. Modesty: minimize self-praise;
5. Agreement: avoid directly disagreements;
6. Sympathy: be sympathetic with others.

These principles, findings and maxims were incorporated while designing Social Plantroid's conversation engine.

Proxemics

Proxemics studies how people inserted in a given sociocultural context use space and interpret the usage of space by others. In other words, it studies the physical and psychological distance people keep from each other in different contexts and, ultimately, how they organize their living and work spaces, such as homes and cities in accordance with such social behavior (proxemic behavior).

Such study consists mainly of three components:

1. Spatial dimensions;
2. Level of interpretation of the spatial dimensions;
3. Physical features of space.

There are four main distances which impact such components ([92]):

1. Intimate zone: immediate physical space surrounding a person, considered to be private or personal space;
2. Personal zone: zone within the reach of an individual, but larger than the privacy sphere, normally used for close friends and family members;
3. Social zone: zone outside of the reach of an individual, normally used for social interactions with acquaintances;
4. Public zone: normally used for public speaking.

Different cultures have different notions on what specific distance ranges constitute such spaces, but, within a given sociocultural framework, researchers can predict how larger are such zones for individuals with statistical accuracy. Since social robots communicate and, sometimes, are able to move, it is necessary to take in to account how the distance between the robot and humans affects comfort levels and how it impacts communication. This particular sub-area of Proxemics is known as Human-robot proxemics, or HRP.

There are a few mathematical models for predicting a comfortable distance between individuals which take into account many parameters, such as how familiar the individuals are, their body

postures, gaze, familiarity of the topic of conversation among many others and, thus, it is important to take such models into account when designing an algorithm for navigation in environments with humans, so robots will move and communicate without disrupting the social environment in which they are inserted. The most prevalent models for inter-human interactions are as follow ([93]):

- Compensation model: suggests an equilibrium between individuals; when one individual approaches or increase eye contact, the other compensates by distancing or decreasing eye-contact;
- Reciprocity model: states the opposite, individuals are more likely to copy the behavior of each other in an interaction;
- Attraction-mediation model: suggests that individuals with initial high levels of attraction will keep close regardless of changes from others, while individuals with low levels of attraction will keep distance despite changes on the behavior of interaction partners;
- Attraction-transformation model: a mix between the compensation and reciprocity models, suggests that the initial attraction between individuals define if they will act accordingly to the compensation (low attraction) or reciprocity models (high attraction).

Research has found that for HRP, how much users like the robot in the first moments of the interaction affects the personal distance from a robot and how much personal information users were willing to share with the robot ([93]). People who have shown higher initial affinity with the robot did not change the distance between themselves and the robot when it increased its eye contact; while people who have shown an initial disliking of the machine increased the distance between them proportionally to the gazing behavior of the robot. Moreover, the last group of people also were less willing to disclose personal information. This shows that, if a person dislikes a robot, it tends to increase the distance between with increased eye contact, supporting the compensation model; but there is also partially support for the attraction transformation model. In the psychological sense, greater evidence shows greater support for the attraction-mediation model ([93]).

Eye-contact is a very important component of intimacy and influences the distance on which individuals feel comfortable while interacting and the proxemic distance tends to reduce when eye-

contact is reduced ([94]). This effect seems to be stronger in men, who have shown, in average, to distance themselves more with increased eye contact than women; and have a higher tendency of keeping a greater distance from robots ([92]).

Age is also an important factor; children tend to prefer interacting with robots in their social zone, while adults have shown to prefer interacting with them in their personal zone ([95]).

Factors such as previous pet ownership and previous robot interaction experiences also influence the distance at which humans feel comfortable interacting with robots; reducing accordingly to increased experience ([92]). However, measuring to what extent people are familiar with pets or robot is very hard, if not impossible, without explicitly asking.

Personal proxemic preferences are not to be considered during conversation, but also during robot navigation; especially because Social Plantroid needs to navigate towards sunlight, shadow and to request help from humans. People, in average, prefer if the robot approaches from the front and stops outside the personal zone. However, such findings are subject to context of what the people are doing, how fast is the robot and many other factors ([92]).

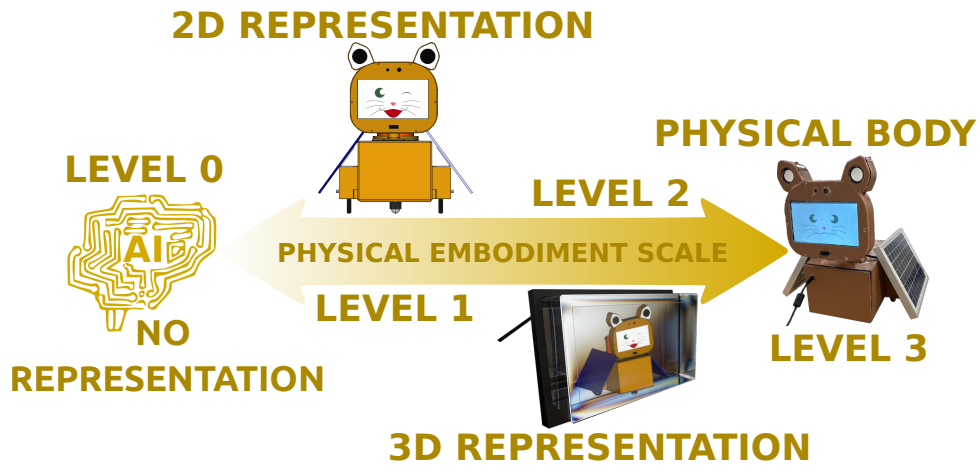


Fig. 2.2: Proposed physical embodiment scale, which goes from no representation (level 0) to a physical body (level 3).

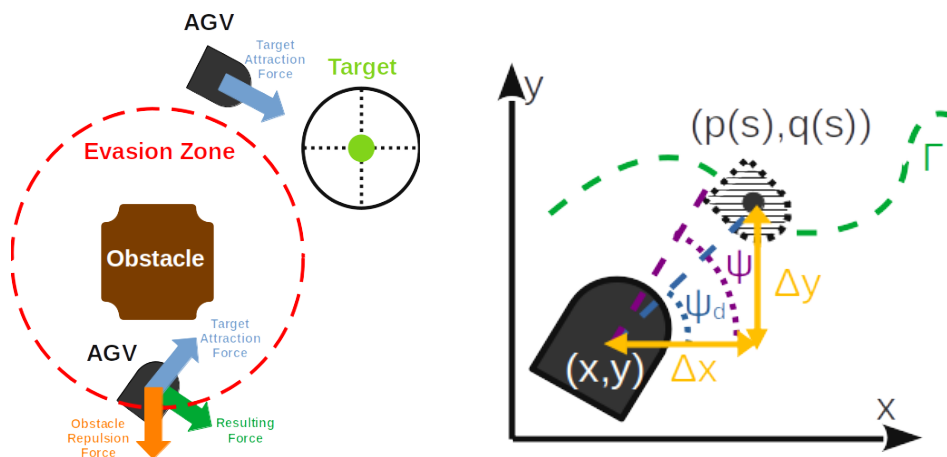


Fig. 2.3: Obstacle and goal point interactions with mobile robots. Fig. 2.4: Virtual Robot Approach main idea.

第3章

Related Works

As for the Background Chapter, the present Chapter introduces works related to the distinct research topics investigated during the elaboration of this thesis. It is also broken into the same subdomains of knowledge: first, Section 3.1 previous research related to phonetics, speech synthesis, multimodal estimation and Pragmatics. After that, Section 3.2 introduces related research works about Embodiment levels, Novelty bias, or preference for novelty. Section 3.3 presents related visual navigation of robots, Proxemics, soil monitoring and plant health estimation research.

3.1 About Prosody in Human-machine Communication

Research on semantic-free utterances is not new, and many different types of semantic-free utterances, such as [7], have been performed, but research on gibberish speech still needs more development. Among the works that used gibberish speech, all of them were based on existing languages, such as Japanese [56] and Dutch and English [12, 96, 97]. Thus, this work is novel in the sense that it presents a language agnostic gibberish speech and analyzes its emotional impact on listeners. It also investigates the effects of prosodic acoustic characteristics of gibberish speech on human impression, but unlike the investigation done in [98], which investigated which prosodic characteristics of gibberish speech better fit different robot morphologies according to children's expectations, it investigates how such characteristics affect adult human impression for a fixed ECA appearance.

In work [97], the authors developed a gibberish generation system based on swapping the vowel nucleus of Dutch and English words to turn them into gibberish, but to avoid ending up with weird sounding words, the authors developed a weighted swapping mechanism according to the probability distribution of each vowel core in English and Dutch. The gibberish generation algorithm developed for the “Talk to Kotaro” experiment deliberately allowed the generation of utterances that did not follow any yule-like phone distribution [99], because if a distribution were chosen, it might cause alienation to speakers of other language families. Moreover, by not following the usual rules, we can study the effects of the violation of such principle on listeners.

In order to analyze what emotions are generated by gibberish speech, the authors of [100] conducted child-robot interaction experiments using an NAO robot equipped with control and behavior modules. The experiments were divided into two trials: one in which the experimental setup was designed to elicit natural emotions in children, and the second in which the setup was designed

to analyze children's perception and response to the gibberish speech of the NAO robot. Similarly, [101] investigates the perceived emotions of Spanish synthetic expressive voices by participants of four Asian nations (Japan, South Korea, Vietnam, and Malaysia), which shows that non-verbal cues are very important in the perception of emotion, but, again, the work does not focus on how the listeners felt.

In [98], the acoustic prosody features were chosen in a Wizard of Oz setup, but several techniques for automatic prosody generation, at least for semantic speech, have been developed. Such techniques can be rule-based [102, 103, 10] or neural network based [104, 105, 106, 107]. In general, rule-based approaches have been superseded by neural prosody selection because manually creating rules to generate appropriate prosody from every possible case is an impossible task. The problem with neural prosody generation is that it depends on existing data from which appropriate prosody for the speech content can be learned; this is not possible for gibberish speech since, by definition, no one speaks gibberish and thus the data is scarce and artificially generated, as in [12]. Thus, this work provides novelty in the sense that it has generated a small dataset that can be used to learn appropriate prosody for IPA-based input text. Moreover, another problem of most neural prosody generation work is that they are tightly coupled to speech synthesis, whereas the proposed architecture is speech synthesizer independent.

Regarding the emotional evaluation of prosodic speech, again, most of the works had semantic speech as their focus [10, 11, 108, 109, 110] and had research participants evaluate their perception of what emotion the generated speech conveyed, rather than how it affected their emotional state. In addition, most of the evaluation was done through subjective post-listening evaluation [110], rather than measuring the immediate response of the participants through their facial expressions and body language, for example through EEG [111]. An exception to these constraints is [100], where they have analyzed the emotion caused by gibberish-speech on children by analyzing the facial expressions and bodily language displayed on video samples.

All of the automatic prosody generation research was conducted for the domain of semantic speech, and most of it focused on learning prosodic patterns from pre-existing audio recordings. Moreover, research studies that develop systems for emotional speech generation have only verified the emotion which listeners perceive on the generated speech, not on how the research subjects themselves felt when listening to the obtained speech, seeking to be perceived as natural speech,

such as [104], which deals with voice only and [105], which also deals with the visual components.

3.2 About Embodiment

Other studies have conducted to investigate how physical embodiment level affects human engagement and perception of ECA [21, 22], and how well users perform certain tasks when interacting with agents of different embodiment levels [23, 24, 25]. However, there is a possibility that such results are due to novelty preference, which may have played an important role in these results. Previous works, even when acknowledging this possibility, have not investigated the effect of novelty preference on participants' impressions and preferences.

Current understanding is that, although not causing significantly better performance of users, higher levels of physical embodiment of ECA seems to cause a higher engagement [22, 23].

3.3 About Social and Agricultural Robotics

The proposed Social Plantroid robot platform, as indicated by its own name, is not the first plant caring robot and surely not the first social robot. This research article is, then, inserted in the very rich research fields of Agricultural robotics and Human-robot interaction (Social robotics); whose related works are listed in Subsections 3.3.1 and 3.3.2, respectively.

3.3.1 Agrobots research

While there is no official definition, consensus is that an agricultural robot is a programmable mechatronic device responsible for performing crop production activities, such as soil preparation, planting seeds, pest control, harvesting *etc* ([112, 28]). While there are many types of agricultural robots, this short review will focus only on mobile ground robots partaking on plant-caring activities in farms, greenhouses and plant-factories, due to their greater similarity with the Social Plantroid and previously developed Plantroids.

[113] developed an open-source autonomous multi-purpose mobile ground robot for plant phenotyping and soil sensing, which is equipped with LiDAR, GPS, stereo camera and a three DoF manipulator arm, which has a soil temperature and moisture sensor on its end effector and a chuck

for attaching different tools. It can be used both in indoors and outdoors applications. In its intended purpose, MARIA does not take care of plants directly, but is responsible for monitoring the quality of the soil and the phenotypic characteristics of crops in order to allow for crop artificial seed selection. However, given its versatile end effector, it might be able to take care of plants.

Since plants require adequate quantities of water in order to grow healthy and water is a precious resource in many regions, precision irrigation is an important research topic for agricultural robotics; thus, many agribots have been developed for that purpose, such as [114, 115, 116, 117, 118, 119]. While Social Plantroid does not irrigate, it closely monitors the moisture level of the soil of the potted plant it carries and requests human for help whenever there is too little water in the soil. The contents of the soil are also very important for the health of the crops and, thus, many different agribots were developed to measure the salinity of the soil ([120]), Ph ([121]), NPK levels ([122, 123]) and other parameters of interesting. The proposed Social Plantroid measures the temperature, moisture, salinity, pH and NPK levels of the soil of the potted plant it carries and stores these values in a SQLite database, which allows it to not only notify humans when corrective action is needed, but to predict when corrective action will be necessary, allowing human caretakers to schedule the procurement of correct fertilizers beforehand.

Research on solar powered agricultural machines and robots is driven by several factors, such as an increasing concern about sustainability, fossil-fuel dependency reduction and difficulty of charging robots on farms and rural areas where electricity might not be easily available ([28, 124]). Moreover, even in greenhouses and urban farms where electricity is easily available, an agribot that does not need to cease operations for charging has great advantages over those who need to. Some examples of solar powered agriculture robots are present in ([26, 125, 126, 127]). The proposed Social Plantroid platform has following sunlight as its primary function, so, recharging the batteries of the robot using sunlight was a natural choice. However, such solar panels only slow down the battery depletion, due to their small size.

The previously developed Plantroid Omni ([26]) and Plantroid mini ([37]) require an external camera to identify sunlit areas and to control their own movements. The novel Social Plantroid, on the other hand, does not require any environment changes or special setup to operate; it uses its own black and white and thermal cameras and photoresistors distributed over its body to detect sunny areas and shadowy areas. Plantroid mini employs an artificial Potential Field-based

navigation method where sunny areas attract robots and obstacles, such as walls and other robots, repels the robot. This navigation algorithm allows a swarm of Plantroid mini to take care of multiple plants, automating a whole plant factory or greenhouse at once. This lack of top-down vision in the new Social Plantroid makes it harder to work in a swarm, since the robot can only detect obstacles in front of it. Moreover, unlike Plantroid Omni, Social Plantroid is not capable of omnidirectional movement, since it is a differential drive robot.

None of the previously mentioned robots has social and speech capabilities, which is a novelty presented by Social Plantroid . The only other agriculture robot which incorporates such aspect is the PotPet ([128]), a pet-like flowerpot robot which has 4 types of sensors: humidity sensor to measure the moisture of the soil of the plant, a light sensor to detect sunlight, motion sensors to detect human presence and ultrasonic sensors to detect obstacles. Previous Plantroid versions have incorporated the sunlight seeking aspect of the PotPet, but didn't present its social aspect: PotPet approaches humans and uses its movements to convey that the plant requires watering. The proposed Social Plantroid platform incorporates that social aspect and takes it further; it can listen and talk to people, displays facial expressions, is capable of a wider range of body language and recognizes human emotion through video and audio.

3.3.2 Social Robotics

Human-robot interaction (HRI) is a broad and multidisciplinary research field that studies interactions between humans and robots. Since agribots are expected to work side by side with human workers, some HRI research on agricultural robots has been performed ([129, 130, 131]). However, there is little research regarding on social behavior and social robotics, that is, with agrobots that are too social robots. A social robot is designed to perform social interactions with humans to elicit social responses from them, unlike robots that are designed to exclusively to perform an external mechanical task ([132, 133]). The idea of developing plant-based social robots is not novel and that is to be expected: humans having been interacting with plants since the dawn of time. Taking care of plants positively impacts human quality of life in a physical and psychological sense and, thus, no developed social plant robot completely takes care of plants, because this would negate the need of humans interacting with the plant and, thus, with the robot itself. The proposed Social Plantroid is not different in that regard, it only monitors and warns users about

the plant's condition and it is up to them to make care-taking decisions and to take necessary corrective actions. A vast corpus of research has been performed about social robots but, for the sake of brevity, this Subsection will dedicate itself on listing only previously developed open-source social robots and plant social robots. A problem in social robotics research is a lack of available customizable robots for meeting different research necessities and, thus, building robots becomes a very costly and time-consuming necessity, whenever commercially available social robots such as PARO ([134]), NAO ([135]) and Pepper ([136]) cannot meet the researcher's needs ([137]). In order to solve such problems, open-source social robot kits such as the OPSORO's Grid System [138] and TJBOT ([139]) were developed; as well as open source software frameworks, like OPSORO [138] and HARMONI ([140]).

Since humans tend to display a higher degree of empathy towards anthropomorphic robots ([141]), many Social Robot platforms have a high degree of anthropomorphism. One of such robots is Ono, a humanoid anthropomorphic social robot which is modular, easy to build (do-it-yourself) and is capable of a wide range of facial expressions – Ono's face has 13 degrees of freedom. Another anthropomorphic open-source social robot is CASTOR [142], developed as a platform for therapy of children with autism spectrum disorder, integrating soft actuators and compliant mechanisms for safe interaction. It 14 DoF, is capable of a wide range of facial expressions, of moving its head and arms and of answering with sentences, sounds and movements. Woody ([143]), besides having an animal-like head, still retains a somewhat anthropomorphic shape, since it is described as an open-source humanoid torso robot. It possesses two arms with five degrees of freedom and a two DoF neck supporting a head with two movable eyebrows. The previously mentioned platforms, however, are not able of moving around, while Nelson ([144]), a low-cost open source social robot for education, is capable of doing so by having an iRobot Create platform at its base. It possesses a three DoF neck, seven DoF face, allowing for many facial expressions and two arms with four DoF each. The aforementioned robots are used mainly for educational and therapeutic purposes, which is somewhat outside of Social Plantroid intended social purposes. However, it can still serve educational purposes, teaching how to take care of plants or teach about plant species in educational gardens, which are powerful educational tools about sustainability, ecology and taking care of nature ([145]).

However, humans also have a tendency to hold anthropomorphic robots to a higher standard

expecting more knowledge a more competent behavior ([141]). Thus, not all social robots should be anthropomorphic, such as Romibo ([146]) a low-cost open-source do-it-yourself and highly customizable robot for motivation, therapy and education. Another non-anthropomorphic open-source robot is Blossom ([147]), which has a floating head platform which is actuated by means of cables and servomotors and is able to rotate its base. The differentiation factor of Blossom is that its external body is soft and handcrafted, allowing the researchers to customize its appearance and expressiveness to a very high degree. The appearance of Social Plantroid was chosen to be pet-like, because it relies heavily on users for taking care of the plant whose health it monitors. Moreover, since the proposed platform can communicate with sounds or with semantic-free utterances ([148]), an animal appearance was deemed to be more appropriate.

Regarding other plant-based social robots, PotPet ([128]) is the closest to Plantroid for mixing both the social aspects and plant-caring; but it is only capable of non-verbal communication. Another social plant robot, flona([149]) communicates with users by moving whenever an ultrasonic sensor detects hand movements. The movements of the plant are generated by strings attached to the body of the plant, which are pulled by stepper-motors. In [150], a cyborg consisting of an iRobot Create base, a plant, light sensors and ultrasonic sensors is used in an artistic installation, where a probabilistic planning algorithm was used to schedule the actions of the cyborg in a way where all physiological needs (water and sunlight) of the plant were satisfied and the iCreate base did not run out of battery, while presenting interesting movement patterns. The objective of the artistic installation was to make the public question the role of plants in society, since they normally are static. All of the plant robots presented above interact with humans only through non-verbal communication and do not possess any capacity of estimating the emotional state of users. Social Plantroid, on the other hand, not only can interact through the movements of its base and its head, but is capable of synthesizing and understanding speech, being able to convey the necessities of the plant clearly, to hold conversation and following users commands.

The proposed open-source social plant-caring robot platform Social Plantroid is, then, a multi-purpose robot that incorporates in a single machine capabilities that were spread through multiple platforms or were not available at all in open-source robots. This way, it has the potential of becoming an useful research platform both for agrobot and social robotics research fields. Additionally, since the project is open-source, researchers are not only free, but invited, to modify the

robot to meet the requirements of their own research topics.

3.4 About Visual Robot Navigation

Since visual robot navigation is a problem of great interest, there is a large corpus of previous works dedicated to solving many different aspects of it. Since this work uses monocular images and neural network to achieve end-to-end navigation without reinforcement learning, this literature review will list only works related to these sub-domains of the visual navigation research.

In [151], a complete visual robot navigation platform consisting of a ResNet50 neural network is proposed; which is trained with navigation trajectories generate by a Model Predictive Control navigation algorithm, whose simulations were ran in a custom simulator HumANav. The present work, however, does not require a custom simulator, using Gazebo and readily available models. Moreover, by using the simpler artificial potential field method to generate the trajectories for the training data, it speeds up implementation.

In [52], a Faster Region-based Convolutional Neural Network (Faster R-CNN) based architecture is trained to detect tree trunks from monocular camera images, while a control strategy which uses the height of the trees to estimate the distance to the obstacles and uses the distance between trees to determine the widest free space in order to safely navigate.

In [152], a CNN is used to predict potential paths for SLAM using uncalibrated 360° spherical images. In [153], a CNN classifier is used for autonomous robot navigation, deciding from monocular RGB images if the robot should move ahead, turn left or turn right in order to avoid obstacles and reach its destination. To train such a classifier, data generated by human operators controlling the robot in a real environment was used. These papers require human operators to generate the training data, which, in this work, is achieved automatically by running simulations in Gazebo.

More closely related, work [154] developed an end-to-end method for training CNN for autonomous navigation of mobile robots using only a RGB-D camera, which is more complex and expensive than the simple gray-scale camera used in this work; and the labeling of the data was not done through the APF method.

The architecture presented in this work is simpler than the ones presented by [155] and [151] [153].

第4章

Obtaining data: the Talk to Kotaro Experiment

In order to obtain the data set necessary for developing a system that allows ECA to estimate the impact their utterances has on listeners, an experiment where volunteers would hold conversations with an embodied conversational agent that speaks Gibberish Speech only was proposed. Again, that was necessary due to the nonexistence of readily available data sets from which good prosody patterns for gibberish speech could be learned from.

Moreover, the COVID-19 pandemic had impacted life in every aspect; and research was not an exception. Economic crisis, silicon crisis, budget reduction, activity-restricting policies, conferences being postponed and, worst of all, a grim death toll. *In-person* experiments became impossible during the period of time when this research was being carried out because it would be very difficult — and irresponsible — to gather volunteers in person at Mizuuchi lab. Thus, the experiment had to be held online, like many other activities. The web-based crowd-sourcing platform was then created to allow circumventing such restrictions. This chapter is dedicated to explain the proposed experiment in an in-depth manner, as well as describing the development process of the platform and of its implementation. The structure of this chapter is as follows:

The experiment consisted of having volunteers hold turn-based conversations with an avatar of the robot Kotaro [14], which answers with semantic-free utterances (SFU) [55] constructed using International Phonetic Alphabet [65] (IPA) symbols. The goal of the experiment was to record what volunteers are saying and their facial expressions while listening to the response of the avatar, in order to estimate their impression regarding the phone and prosody choices of the system.

This data will be used to generate a phone-prosody embedding for the robotic utterances, clustering phones and prosody according to the estimated human impression on them; that is, phones said with certain prosody patterns will be close to other phones and patterns which generate similar impression on humans, in a word2vec [156] fashion. Other works have performed phoneme-embedding in the context of verifying if the phoneme distribution in languages implied similar meaning between words composed by such phonemes [157], emotion recognition of speech [158] and automatic speech recognition (ASR) [159], but without considering prosody parameters and outside of the HRI context.

Crowdsourcing data from all over the world is essential in that context, because we intend on analyzing how people from different cultures react to SFU. The initial hypothesis is that even if there are different impressions, there may be a common baseline, in similar fashion to the Bouba-

Kiki effect [15]. The objective of the experiment is to gather data that will allow us to: (i) test the hypothesis that there is some common baseline on how people from different cultures react to different phones and prosody patterns and (ii) if there is a baseline, to develop a human impression prediction module using the crowdsourced data.

However, there was no appropriate platform that allowed volunteers to talk with a robot in their web browsers while audio and video from such interactions are securely streamed to our lab servers. The closest tools we were able to find didn't have all the necessary features [16], were too game-like [17], introducing many other factors that could impact the impression of volunteers, or required a VR Headsets [18], which make impossible to record user facial expressions and not web-based [19]. This way, we decided to develop our own solution and make it open-source, so it can help other HRI researchers hold their own experiments online, saving time and implementation costs. It was designed in such a way that others can easily use it and modify it according to their needs. Moreover, this tool is helpful not only during times of crisis. Given that crowdsourcing information from all over the world will make obtained data sets more diverse and, thus, research will be more robust.

4.0.1 Talk to Kotaro: an web crowdsourcing experiment

4.1 Experiment Description

To better understand how people respond to different phonetic and prosodic choices in gibberish speech, the web-based crowdsourcing experiment “Talk to Kotaro” was conducted between 2021/10/1 to 2023/03/31 and has been approved by Tokyo University of Agriculture and Technology Ethics Committee (approval number 210801-0321 and experiment extension request approval number 220306-0321). All participants had to read an online consent form, the experiment instructions and had to click a consent button, which was deemed as an acceptable means of obtaining consent by University of Agriculture and Technology Ethics Committee. It was a Human-Computer Interaction experiment whose objective is to collect audio and video data to allow us to better understand the impact of phone and prosody choices for synthesized speech on human impression. The aforementioned data consists of audio recordings of what volunteers tell Kotaro, video of the facial expressions made while listening to Kotaro's GS, the text information that users

provide while registering their profile for the experiment and their answers to a 10-question long Likert scale [60] questionnaire (which is optional). Profile information will be used to better understand the impact (if there is any) of region of origin, spoken languages and the cross-cultural experiences have on human impression of synthesized Gibberish speech.

4.1.1 Structure of the experiment

First time volunteers need to read the consent form, experiment instructions and, if they agree on participating, create a profile, for which they need to provide relevant information regarding their backgrounds which might influence their impression on GS. Such information will be detailed in Subsection 4.1.3.

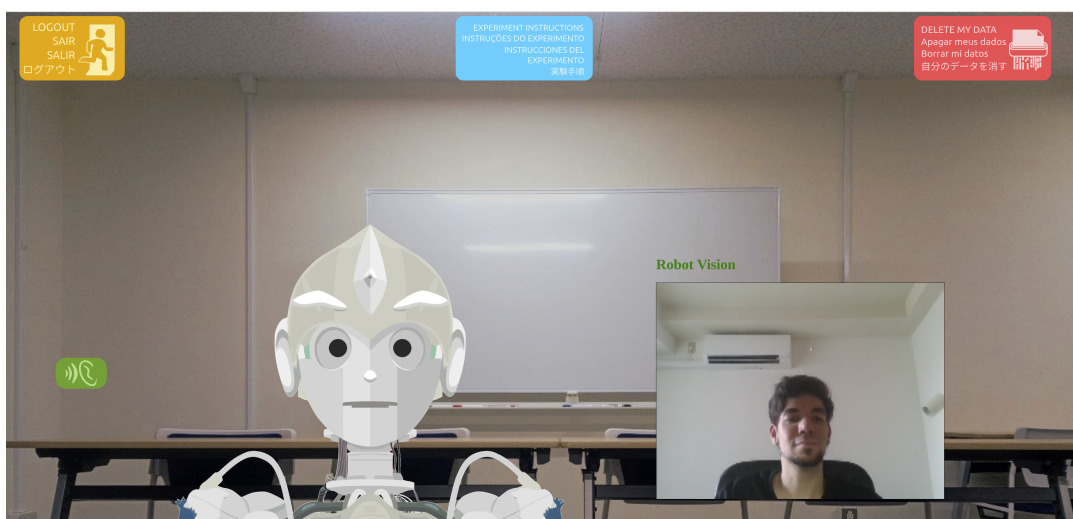


Fig. 4.1: Talk to Kotaro experiment: crowdsourcing human impression information online.

After creating a profile, users need to login and start the experiment. The experiment screen, shown in Figure 4.1, is a web page where volunteers are free to have a turn-based conversation with Kotaro for as long as they want. In order to start chatting, test subjects must press the green button, which will become blue. The web page will record what the participant is saying until the button is pressed again. That audio is sent over the network for posterior emotional analysis. On the server side, GS will be generated using an algorithm described in Subsection 4.1.2, and sent over the web. Kotaro moved its mouth while the volunteer's web browser plays the utterance. While Kotaro moves its mouth, the facial expressions of volunteers will be recorded by their webcams; and send

to the server. This is done in order to allow researchers to run emotion estimation algorithms to estimate the volunteers' first impressions based on their facial expressions.

This turn-based conversation dynamic was created to avoid the need for Voice Activity Detection (VAD); and it can be repeated for as long as volunteers have interest in doing so. There was no minimum nor maximum duration for the experiment. The most prolific volunteer contributed 201 conversations, about 23.4% of all the data in the experiment. Whenever participants wish to leave the web-page, they should click on the Logout button and they will be prompted with a Likert-scale questionnaire, which is described in Subsection 4.1.4). The questionnaire is optional in order to avoid volunteers who will click any responses to quickly end the experiment because they are already tired of interacting with Kotaro, and 22 of the 37 participants chose to do so.

The IPA-based gibberish speech spoken by Kotaro was generated using the eSpeak [66] speech synthesizer, which was chosen because it is open source, can receive ASCII-IPA input and allows for controlling the prosody of the generated speech. Algorithm 1, described in Subsection 4.1.2, was used to select the phones to be used in Kotaro's speech. As for the prosody, the three chosen parameters, speed, pitch, and volume, were randomly chosen between 80-450 words per minute (speed), 10%-200% (volume), and 0-99 (arbitrary unit, pitch). Some participants reported a feeling of alienation when the ECA suddenly changed its voice pitch, making them feel like they were not talking to the same person.

4.1.2 Gibberish speech generation algorithm

To create Kotaro's gibberish, an algorithm, originally described in [13], draws vowels and consonants from the IPA table to randomly generate Kotaro's responses. The International Phonetic Alphabet, IPA, is a phonetic notation system created by the International Phonetic Association in the 19th century to provide a standardized way of representing speech sounds in different languages in written form [65]. It can represent various aspects of the lexical and prosodic sounds of human speech; phones, intonation, and pauses. Other non-speech sounds, such as clicks, grits, and lisping, are represented by an extended set of symbols. There are two basic sets of symbols: letters and diacritics, which are showcased in Figure 4.2.

While designing the experiment, it was necessary to decide how to attribute a probability of selection for each phone, because every language has its own Yule-like [99, 161] phoneme dis-

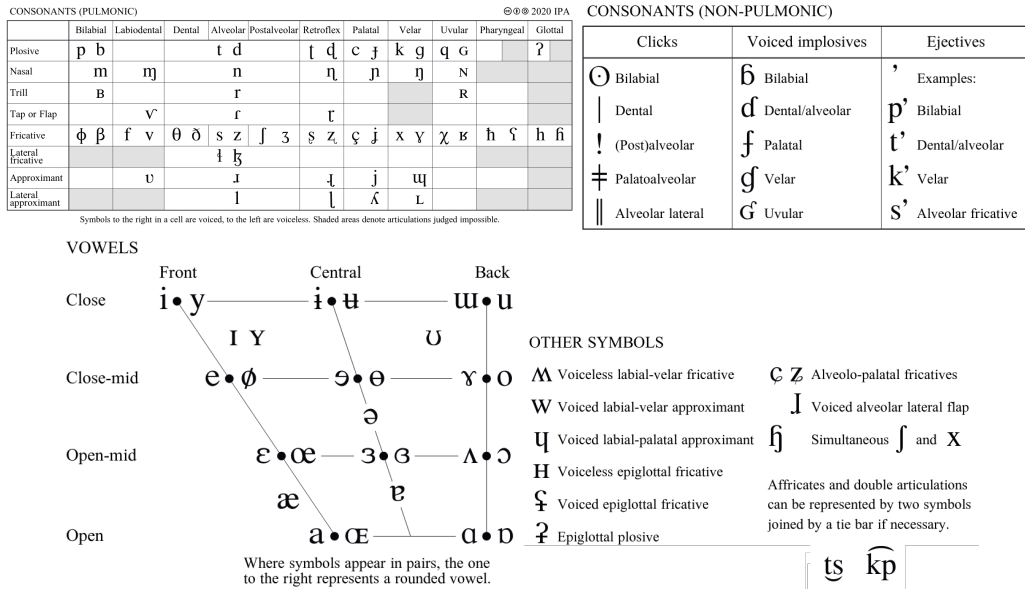


Fig. 4.2: Consonant, vowel and other symbols IPA charts, taken from [160].

tribution (which implies a similar phone distribution). However, if we did select a non-flat distribution, the GS might resemble an existing language, which would bias the enjoyment of the Talk to Kotaro experience towards speakers of certain families of languages. However, by not choosing a Yule-distribution, we risked making the GS too strange and impossible to be enjoyed by any volunteers. However, since there is no work supporting that Gibberish Speech whose phone choices do not follow a Yule distribution can cause listeners to be estranged, we have decided to have a “naive” phone probability distribution, that is, every IPA symbol has the same probability of being picked. In this sense, this is another novelty provided by this research, since other other research use a phoneme distribution close to an existing language, e.g. English and Dutch [12] and Japanese [56]. Following such design principles, Algorithm 1 was proposed, whose pseudo-code 1 describes how Kotaro’s utterances were randomly generated during the experiment.

To better understand Algorithm 1, it is necessary to define the function *choice(l)*, which randomly chooses an element belonging to the list *l*. At the beginning of the routine, the number of iterations for generating the utterance is randomly chosen between 1 and 10, an arbitrary maximum chosen by the researchers to avoid very long utterances and to avoid very long delays between a volunteer finishing speaking and Kotaro responding. The utterance starts as an empty string, which

Algorithm 1 IPA Giberish Speech generation algorithm

```

1: procedure GENERATE GIBBERISH
2:    $max_{iter} \leftarrow choice([1, \dots, 10])$ 
3:    $counter_{iter} \leftarrow 0$  ▷ iteration counter.
4:    $utterance \leftarrow ""$  ▷ gibberish speech utteranc, starts empty.
5:    $IPA_v$  ▷ list of all IPA vowels.
6:    $IPA_c$  ▷ list of all IPA consonants.
7:    $IPA_o$  ▷ list of all IPA other symbols.
8:   while  $counter_{iter} < max_{iter}$  do
9:      $chunk \leftarrow choice(choice([IPA_v, IPA_c]))$ 
10:     $chunk \leftarrow chunk + choice(choice([IPA_v, IPA_c, IPA_o, ""]))$ 
11:    if  $len(chunk) > 1$  then
12:      if  $chunk[0] \in IPA_c \wedge chunk[1] \in IPA_c$  then
13:         $chunk \leftarrow chunk + choice(IPA_v)$ 
14:      else if  $chunk[1] \in IPA_o$  then
15:         $chunk \leftarrow choice(choice([IPA_v, IPA_c]))$ 
16:      else
17:         $chunk \leftarrow chunk + choice(choice([IPA_v, IPA_c, [], [], [], [], [], []]))$ 
18:     $utterance \leftarrow utterance + chunk$ 
19:     $counter_{iter} \leftarrow counter_{iter} + 1$ 
return  $utterance$ 

```

gets chunks of one or more IPA symbols in each iteration. There is a 50% chance that a chunk will start as a vowel and a 50% chance that it will be a consonant symbol. All symbols in each list have the same chance of being chosen by the *choice* function. After that, there is a 75% chance that a second symbol will be added (vowels, consonants, and other symbols all have a 25% chance), and a 25% chance that nothing else will be added to the chunk. If a second symbol is chosen, and both the first and second are consonants, a third symbol from the vowel list is added. If the second symbol chosen is another symbol, there is a 50% chance that a vowel will be added, and a 50% chance that a consonant will be added instead. Otherwise, there is a 12.5% chance that a vowel will be added, and a 12.5% chance that a consonant will be added. The remaining probability is that nothing will be added. At the end of the iteration, the chunk is added to the utterance and the iteration counter is incremented.

Note that when receiving ASCII-IPA input, eSpeak will skip unpronounceable sounds if there is a space between each chunk, i.e. it will just speak the next one. An example of an utterance generated by this algorithm is: ionu'i:inə'ə.

With the contents of the GS chosen, it is now necessary to select the prosody parameters for the utterance. For the sake of simplicity, a single set of prosody parameters are used during the whole utterance, because using Speech Synthesis Markup Language tags to give multiple prosody parameters in a single utterance would make the analysis of the results even harder. For the speed, a value between 80 words per minute and 450 words per minute is randomly chosen. For the pitch, a value between 0 and 99 (no unit is provided in the documentation) is randomly chosen. Finally a value between 10% and 200% is chosen for the loudness. Such values are the lower and upper limits of eSpeak prosody parameters. The only exception is the lower limit of loudness, which was chosen as 10% in order to generate low-volume, but not completely silent, utterances. A representation of the prosody of generated gibberish speech can be seen in Figure 4.3.

4.1.3 Profile Information

In order to verify if the cultural background of volunteers influences their impression on GS and to investigate how, during the profile creation step of the experiment the following information is necessary:

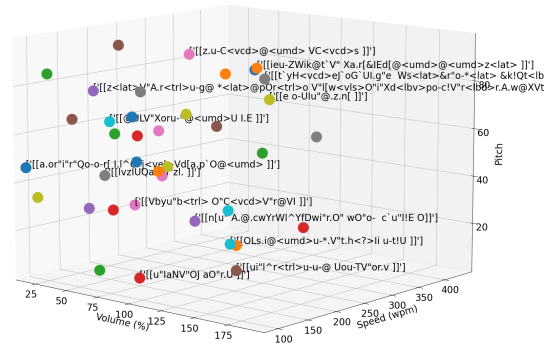


Fig. 4.3: Prosody Space of generated gibberish speech

- ID;
- Password;
- Age;
- Gender;
- Country/Region of Origin;
- Mother language;
- Other languages you speak;
- If you live or have lived abroad, write where;
- Years living abroad.

ID and password are necessary in order to allow a volunteer to send data and to have it securely cryptographed, while correlating it with his or her profile information.

Age and Gender are asked in order to enable us to investigate if some age brackets have distinct tastes regarding GS when compared to others, the same goes for gender.

Country or origin, Mother Language, Other Languages you speak, if the volunteer has lived abroad and for how long – are asked to try and investigate the impact of the cultural background of volunteers in their reactions to GS phone and prosody choice.

4.1.4 Likert Scale Questionnaire

Likert scale questionnaires are a tool for measuring overall attitudes toward a topic. They consist of prompts, statements about the topic being studied, to which respondents choose their level of agreement, ranging from strongly agree to strongly disagree. The number of prompts and possible responses is not predetermined; researchers must use as many as they need, keeping in mind that increased precision may be offset by increased burden on research subjects. However, the most traditionally used scales have either five or seven responses. It is also possible to remove the neutral option, that is, to have a pair of possible levels of agreement, to prevent respondents from over-relying on neutral responses as a socially acceptable stance. This paper uses the traditional 5-point scale and 10-point prompts to avoid tiring respondents.

The Likert scale questionnaire used in the Talk to Kotaro experiment uses a classic five-point format, i.e., respondents can choose their level of agreement with a prompt between 1—strongly disagree, 2—disagree, 3—neutral, 4—agree, and 5—strongly agree.

This decision was made so that respondents would not have to think too much while answering a questionnaire that they could simply exit by closing a tab on their web browser. However, not making the questionnaire mandatory was a design choice to prevent participants who were already tired from the experiment from randomly clicking through the answers to end their participation as quickly as possible. While this risk could not be completely avoided, as participants completed the questionnaire unsupervised, it was a way to reduce this possibility.

The questionnaire was designed to measure volunteers' enjoyment of the statements Kotaro responded to them with, and to measure what factors were most relevant to that impression. The prompts shown are as follows:

- (P_1) Talking with the robot avatar was interesting;
- (P_2) Variation of the speech characteristics made conversation more natural;
- (P_3) Some randomly generated words are less pleasant than others;
- (P_4) Some speech characteristics, such as speed, loudness or pitch influence more than others;
- (P_5) Different random words didn't have an impact on your enjoyment;

(P_6) You felt that the robot was answering your speech accordingly;

(P_7) Longer phrases were more interesting;

(P_8) The turn-based conversation felt unnatural;

(P_9) Foreign sounding phrases were more interesting;

(P_{10}) The robot seemed to be intelligent.

4.2 Platform Description

The platform hosting the “Talk to Kotaro” HRI experiment was built with modularity and easiness of modification in mind. There are three key components to the platform: the web-page templates (the front-end), the server application (back-end) and the memory (file storage and database). Only the server application is containerized inside a Docker container, the web-page templates and the memory are bind-mounted to the Docker container, so all stored data is preserved when the Docker container is destroyed; and the web-page templates can be easily edited without the need to rebuild the docker image. The overall Platform structure is shown in Figure 4.4 and each component will be described in detail in the following Subsections.

4.2.1 Server Side - Server Application

This is the main module of the platform, which is responsible for: Login information verification, serving web-pages, generating and sending Kotaro’s Semantic Free Utterances, audio and video capture, cryptography and storage. It consists of the following python scripts:

- *flaskapp.py*: the main file, responsible for serving the web-pages and calling every back-end functions;
- *login.py*: contains functions that verify if an ID already exists, login information verification and login creation;
- *encryption.py*: contains all encryption and decryption functions, uses AES-256;

- *mailer.py*: contains the functions which enable the application to send profile delete request emails;
- *emotion_estimator.py*: contains the CNN Classifier which estimates the impression of volunteers;
- *utils.py*: contains the gibberish speech generation function, data URI to cv2_img and other useful image manipulation functions.

Besides those files which are essential for running the experiment, there is the dockerfile, which specifies all necessary programs and python libraries for the docker environment on which the application will run; that is, it specifies that it should run an Ubuntu 20.04 image, install espeak, flask, python-opencv, *etc.* Currently, in order to reduce the server load, the emotion estimation function is not being performed as image frames arrive, but that might be easily enabled.

This module of the server application also contains all static files which are served to the client, such as Kotaro's avatar images, Javascript scripts responsible for controlling how the avatar behaves, capturing audio and image *etc.*

4.2.2 Server Side - Memory

The memory module consists of a SQLite database and two folders which contains, respectively, the cryptographed audio and video files. The SQLite database contains the following tables:

- *login*: stores the ID the hashed password, and the salt for the cryptography algorithm;
- *info*: stores all profile information of volunteers;
- *likert*: stores Likert scale questionnaire responses;
- *analysis*: stores Kotaro's utterances and the associated volunteer speech and reaction video files names.

4.2.3 Server Side - Web-page templates

The web page template module contains all HTML, CSS and some of the Javascript files necessary for the client side. It contains a landing page, which explains the experiment, a login page,

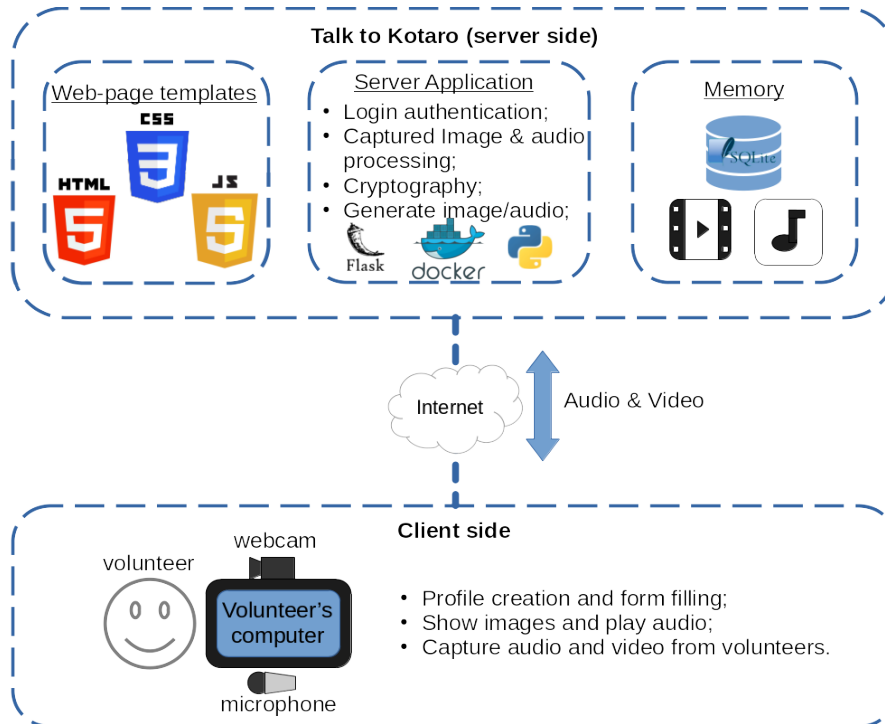


Fig. 4.4: Talk to Kotaro Platform Structure.

a consent form and experiment instructions page, the main page of the experiment, a data deletion page, a page where legal guardians can request the deletion of data belonging to minors and the Likert scale questionnaire page.

Audio recordings of the volunteers conversations with Kotaro are sent over the web using the Recorderjs [162] library and the video of volunteers' facial expressions are captured by their webcams, displayed on a canvas in the experiment webpage and sent over the web using a XMLHttpRequest, frame by frame.

第5章

Analysis of obtained data

In order to analyze the data obtained through the Talk to Kotaro experiment, it was necessary to label the obtained data in terms of valence and arousal, in order to obtain the emotional change caused by the distinct phone and prosody choices for the gibberish speech of Kotaro. To do so, it was necessary to implement several neural networks for analyzing the emotion of the audiovisual data provided by the volunteers, which are described in this chapter.

5.1 Neural Network Architectures for emotion analysis

The most important data provided by the volunteers of the “Talk to Kotaro” experiment consisted of video recordings of the facial expressions of participants while listening to the gibberish speech responses and audio recordings of what participants told the conversational agent. The idea behind recording both audio and video was to help gauge the emotional state of volunteers before Kotaro’s answer, from the audio, and understand how the emotions of the volunteers have changed from the facial expressions displayed in the recorded videos. To achieve such a goal, three different artificial neural network architectures were employed. Two neural architectures, described in Subsection 5.1.1, are used for estimating the emotion of volunteers from their facial expressions; and one, described in Subsection 5.1.2 is used for sentiment classification of the audio samples.

After possessing labels in terms of valence and arousal for the reaction displayed by volunteers after listening to distinct phones and gibberish speech patterns, we investigate the relationship between the aforementioned parameters of Kotaro’s Gibberish Speech and the impression of volunteers through Neural Network architectures described in Subsection 5.1.3.

5.1.1 Emotion estimation from video

In order to obtain the impression created on the volunteers by the GS utterances, two different neural network architectures, VGG-16 and ResNet18 (inspired by the architecture proposed in [163]), were used to estimate the volunteers’ valence and arousal, respectively, from the videos of their facial expressions. This hybrid system was chosen because VGG-16 performed better than ResNet18 for valence, while ResNet18 performed better for arousal. Both networks were trained on the AffecNet data set [164]. This method of engagement and preference estimation was chosen because it is not an invasive method, does not require very expensive additional hardware for

volunteers (given most laptop computers, tablet computers and smartphones have front cameras nowadays), and it does not pause the experiment, allowing one to capture the immediate emotional change caused by the speech sound. The decision for capturing the immediate reaction stems from previous research findings that the candid reaction of research subjects differs substantially from their opinion after being given some time to think and rationalize their own feelings and opinions about an experiment [165, 166]. However, since it is also important to know the attitude of volunteers towards Kotaro’s gibberish speech, towards Kotaro and the experiment itself after having some time to think, this approach is coupled with the Likert scale questionnaire proposed in Section 4.1.4.

However, since the aforementioned neural networks estimate human emotions from still images, and the collected data consists of video samples, it was necessary to choose a metric capable of representing the impact of the gibberish speech on the listener. Thus, it is necessary to obtain the initial emotional state E_t of the subject and the emotional state E_{t+1} after listening to the utterance.

The chosen metric is then the difference between the emotion estimation from the initial (just before Kotaro starts speaking) and the last frames of each video sample. This metric is called $\vec{\delta}_E = (\delta_v, \delta_a)$, where δ_v, δ_a is the change in displayed valence and arousal. It is possible to see how a few utterances (randomly selected from the data set) have affected the subjects in the valence-arousal space shown in Figure 5.1.

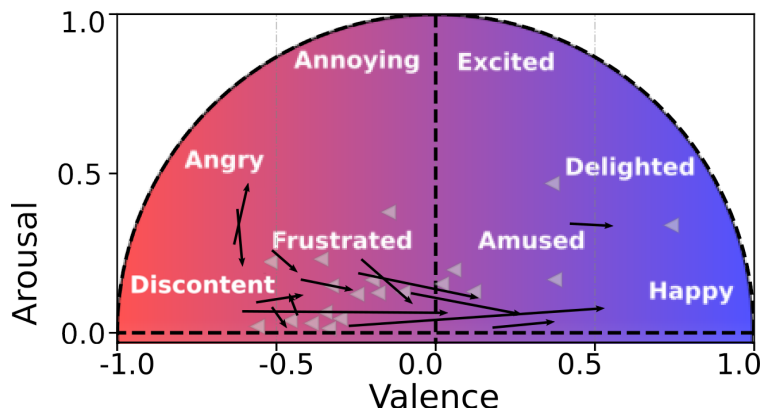


Fig. 5.1: $\vec{\delta}_E$ represented in the valence–arousal emotion space, where arrows indicate the valence–arousal change and gray triangles denote an utterance that caused no visible emotional impression.

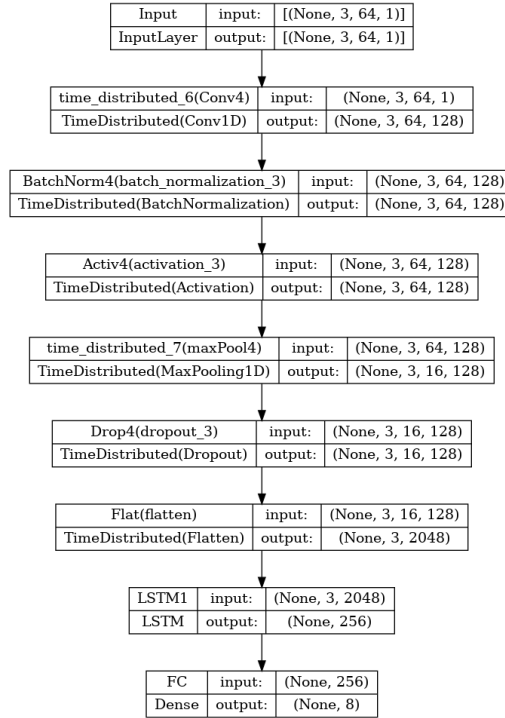


Fig. 5.2: Architecture of the LSTM-based Neural Network used for analyzing the sentiment of the voice recordings of participants.

5.1.2 Sentiment analysis of recorded speech

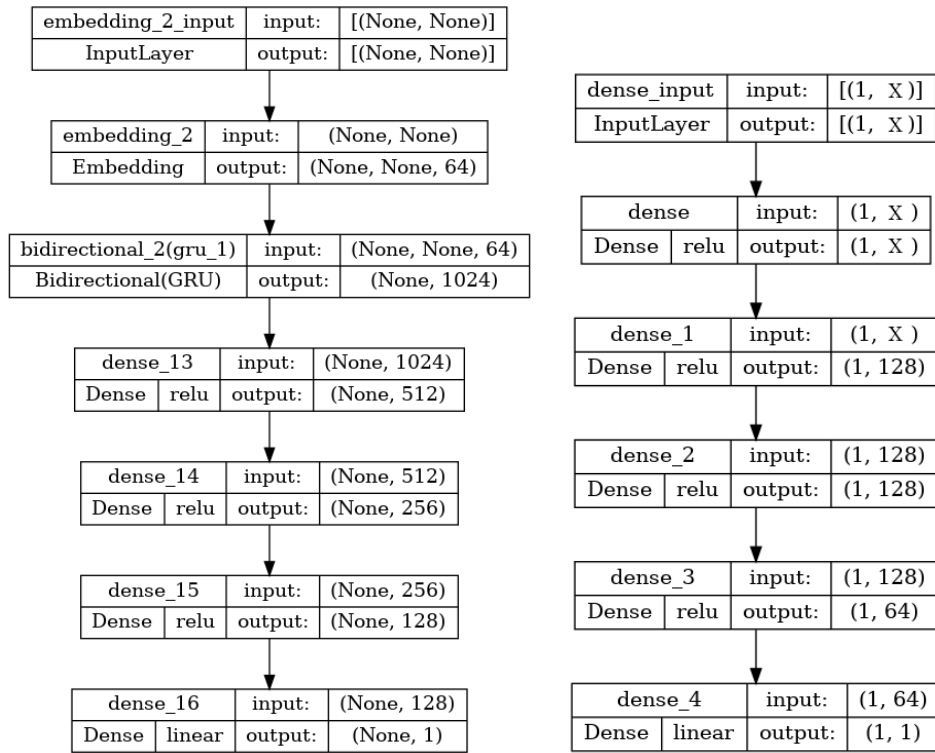
To make the predictions of the volunteers' initial emotional state just before listening to Kotaro's responses more accurate, the Talk to Kotaro web platform recorded what the volunteers said to Kotaro, which allowed us to perform sentiment analysis on the recorded audio samples. However, the initial problem is that the authors could not find a dataset for human speech whose sentiment labels were in terms of valence and arousal, only categorical labels. The chosen datasets were the audio datasets TESS [167], RAVDESS [168] and SAVEE [169], whose samples were labeled with one of the following 7 emotions: anger, disgust, fear, happiness, neutral, sadness and surprise. We extracted the main features of the available audio data using Mel Frequency Cepstral Coefficients (hereafter called MFCC) and used the obtained information to train an LSTM-based neural network, whose architecture is shown in Figure 5.2, which was then used to verify the accuracy of the participants' initial emotional state prediction, at least in qualitative terms, since the labels are categorical.

5.1.3 Gibberish Speech Impression Prediction System architecture

To better understand the effect of phones on participants' impressions, a neural network called GRU_{phones} was built, consisting of an embedding layer with 64 outputs, followed by a bidirectional Gated Recurrent Unit (GRU) layer with 512 units, which is then followed by 4 fully connected layers with 512, 256, 128, and 1 neuron. All connected layers use ReLu as activation function, except the last one, which is linear (architecture shown in Figure 5.3 (a)). The proposed neural network was able to learn an embedding for each of the 71 IPA symbols used by Kotaro (some symbols were not used because not enough utterances were generated). The neural network was trained with the data from the experiment, taking the tokenized IPA symbols as input and outputting the predicted valence or arousal.

Besides the analysis performed by Stuart–Kendall's τ_C correlation coefficient, another way to learn the correlation between the acoustic prosodic parameters is to use a neural network that receives as input a vector containing the speed, volume, and pitch of a given utterance and predicts the subjects' impression. However, only using the prosody information did not yield good results, and by adding the profile information encoded together with the prosody parameters into a 1×80 vector, it was used as input for a neural network called $MLP_{profile+prosody}$, which consists of an input layer of 80 neurons connected to three hidden layers of 128, 128, and 64 neurons each, and ReLu as the activation function (architecture shown in Figure 5.3 (b)). The output layer is a single neuron, and thus, two copies of $MLP_{profile+prosody}$ were trained, one for predicting arousal and another for valence.

Since gibberish speech utterances C , it is necessary to take both aspects into account to make accurate predictions, and thus, we combined both neural networks by averaging their outputs. Other architectures were tested for combining $MLP_{profile+prosody}$ and GRU_{phones} , but the results were not as accurate the ones obtained by averaging the outputs of both pre-trained models. The resulting model is called the **Gibberish Speech Impression Prediction System**, hereby referred to as *GSIP*.



(a) Architecture of the Bi-directional GRU Neural network GRU_{phones} for generating a phone-embedding matrix.

(b) Architecture of neural network $MLP_{profile+prosody}$ and its variations, where X represents the number of columns of the input vector.

Fig. 5.3: Neural networks used for impression prediction in this work.

5.2 Correlation Analysis

With the audio and video recordings properly labeled, it is necessary to calculate the correlation between the Impression caused by the utterances and the acoustic prosody parameters. The Stuart-Kendall's τ_C correlation coefficient test[170] was chosen to perform such analysis, since its assumptions are not as strong as Person's correlation coefficient. That is, it is a non-parametric hypothesis test for statistical dependence which measures rank correlation – how similarly data is organized when sorted by each other quantities. It can be used for discrete data and does not assume a normal distribution, it is, thus, ideal for the present research, since the prosody values used by *espeak* are discrete and the δ_E .

The Stuart-Kendall's τ_C correlation coefficient between two variables is high variables with similar ranks and low for observations with different ranks. It is defined as follows:

$$\tau_C = \frac{2(n_c - n_d)}{n^2 \frac{m-1}{m}} \quad (5.1)$$

5.3 Analysis and results

In this section, we present the data obtained from the “Talk to Kotaro” experiment in fully anonymized form and perform the necessary analysis to verify the influence of phone and prosody choices in gibberish speech. The audio and video recordings cannot be shared because that would violate the privacy of the volunteers, a condition set by Tokyo University of Agriculture and Technology's Ethics Committee. However, a fully anonymized version of the dataset, containing the phones, prosodic parameters of each generated gibberish speech and the results of the emotional analysis performed on audio and video data are available together with its partitions into data set without outliers, training, validation and test data sets are available in the supplementary files of paper [171].

Subsection 5.3.1 presents the profile information of the participants of the experiment, while subsection 5.3.2 presents the results of the emotion analysis performed on the participants' video and the investigation of the correlation between the prosody parameters and the impression on the volunteers. To further improve the emotion estimation from the volunteers' facial expressions before listening to Kotaro's utterances, we performed a sentiment analysis of the participants'

recorded speech, which is presented in Subsection 5.3.3. Subsection 5.3.4 presents and discusses the results of the phone embedding matrices obtained from the experimental data; and the results of the Likert scale questionnaire are discussed in 5.3.5.

5.3.1 Profile of Participants Breakdown

This subsection breaks down the information about the participants of the Talk to Kotaro experiment. Profile information of the volunteers was stored to try to determine how the prosody changes affected each nationality, speakers of certain languages, age groups, etc. The stored information included ID, password, age, gender, country/region of origin, native language, other languages spoken by the volunteer, and if the volunteer lives or has lived abroad (write where and years lived abroad).

The initial goal was to try to find a cross-cultural baseline for human impression for different prosody parameters, a point that will be described in more detail in the Section 5.3.2. The effects of phone choice on human impression are discussed in Section 5.3.2.

Countries with participants are shown in Table 5.1, along with the number of speakers of each language. Initially, 61 participants from 16 countries speaking 17 languages registered, but after removing those who contributed with no data, or contributed only with unusable data (e.g. participated in very dark environments, wore face masks, etc.), only 37 were left. That fact showcases one of the greatest weaknesses of web-based crowdsourcing: data quality varies a lot because participants have different hardware and environment conditions, and might misinterpret instructions without any chance for correction.

Out of the remaining 37 participants, 23 were male and 14 were female. The mean age of the participants was 27.46 years, with a standard deviation of 9.39 years, a median of 25 years, and a mode of 21 years. The youngest participant was 18 years old and the oldest was 55 years old.

Table 5.1: Breakdown of the cultural background of the participants.

Region of Origin	Male	Female	All	Mother Language	Male	Female	All	Total Speakers	Male	Female	All
Japan	9	10	19	Japanese	9	11	20	English	19	15	34
Brazil	6	1	7	Portuguese (Brazil)	6	1	7	Japanese	14	11	25
Malaysia	2	0	2	Mandarin	3	0	3	Portuguese (Brazil)	6	1	7
China	1	0	1	Cantonese	0	1	1	Mandarin	3	0	3
Hong Kong (China)	1	0	1	English	0	1	1	Malaysian	2	0	2
India	0	1	1	Marathi	0	1	1	Arabic	1	0	1
Peru	1	0	1	Spanish	1	0	1	Cantonese	1	0	1
USA	0	1	1	Arabic	1	0	1	Spanish	1	0	1
Bangladesh	1	0	1	Sinhala	1	0	1	Sanskrit	0	1	1
Egypt	1	0	1	Bengali	1	0	1	Korean	0	1	1
Sri Lanka	1	0	1					Sinhala	1	0	1
Undisclosed	0	1	1					Bengali	1	0	1
								Hindi	0	1	1
								Marathi	0	1	1

5.3.2 Impression Estimation from Video and Prosody Correlation

In this subsection, the videos of the volunteers' facial expressions are analyzed and the impression \vec{I}_S , the immediate emotional response to the speech act S , is obtained by the vector $\vec{\delta}_E = (\delta_v, \delta_a)$, which is obtained by subtracting the estimated emotional state of the initial and final frames of the video. In this way, a data set is generated that associates the speech acts and the human impression, allowing us to verify if there is a correlation between the prosody parameters and the impression \vec{I}_S , and to use machine learning to obtain an embedding matrix for the IPA phones. This is achieved by using the VGG-16 and ResNet-18 neural networks described in Section 5.1.

Plotting all obtained $\vec{\delta}_{E,S} = (\delta_{a,S}, \delta_{v,S})$ vectors, the impression caused by each speech S in the valence–arousal space yields Figure 5.4 . However, since there are 734 vectors, it is difficult to visualize the results in valence–arousal space. Plotting the histograms of the results of the analysis, shown in the left and middle histograms of Figure 5.5, shows the results of the analysis performed on the video samples of the participants' reactions to each utterance spoken by Kotaro during the entire experiment. It can be seen that many utterances had little effect on the participants' valence or arousal. However, if we calculate the norm of the $\vec{\delta}_E$ vector for each speech act, we obtain the right histogram in Figure 5.5, which shows that many utterances caused little to no change in the emotional state of the listeners, but most still made an impression. The set of all $\|\vec{\delta}_E\|$ has a mean of 0.124 and a standard deviation of 0.135.

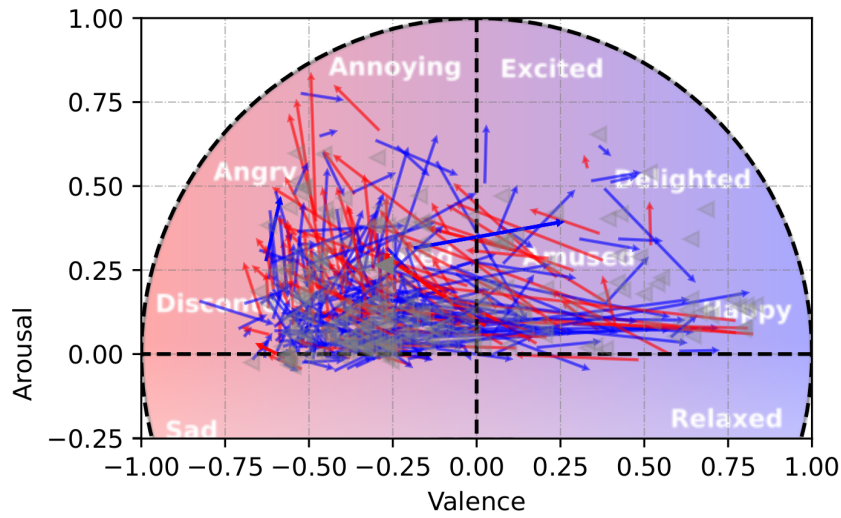


Fig. 5.4: Emotional state changes caused by every utterance in the “Talk to Kotaro” experiment in the valence–arousal space, where a blue arrow denotes a positive change in valence, a red one denotes negative valence change, and a grey triangle denotes no visible emotional change.

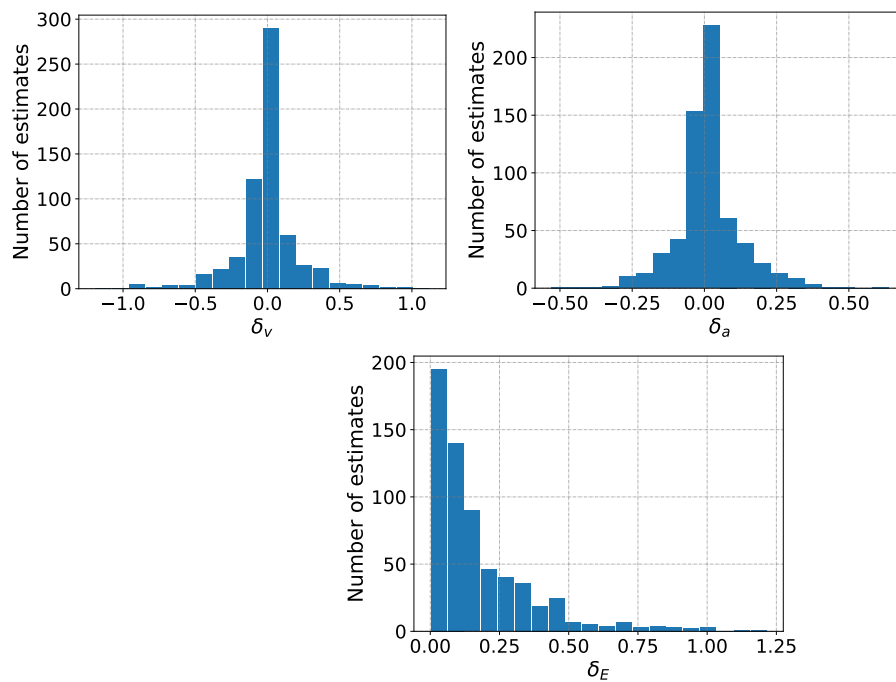


Fig. 5.5: Histograms of δ_v (top left) and δ_a (top right) and $\|\vec{\delta}_E\|$ (bottom) for every utterance generated in the “Talk to Kotaro” experiment.

Since Russell’s two-dimensional model of valence and arousal is defined over $\{v \in \mathbb{R} \mid -1 \leq v \leq$

1} and $\{a \in \mathbb{R} \mid -1 \leq a \leq 1\}$, where v and a are valence and arousal, respectively, the impression space is defined over $\{\delta_v \in \mathbb{R} \mid -2 \leq \delta_v \leq 2\}$ and $\{\delta_a \in \mathbb{R} \mid -2 \leq \delta_a \leq 2\}$. Thus, we can obtain another representation for all the impressions caused by Kotaro's utterances, shown in Figure 5.6, where outlier impressions are highlighted in red. Since the obtained impressions consist of two variables, we used the Mahalanobis distance metric [172], which measures the distance between a point and a distribution, to determine which emotion changes were outliers, with a Mahalanobis distance threshold of 3.

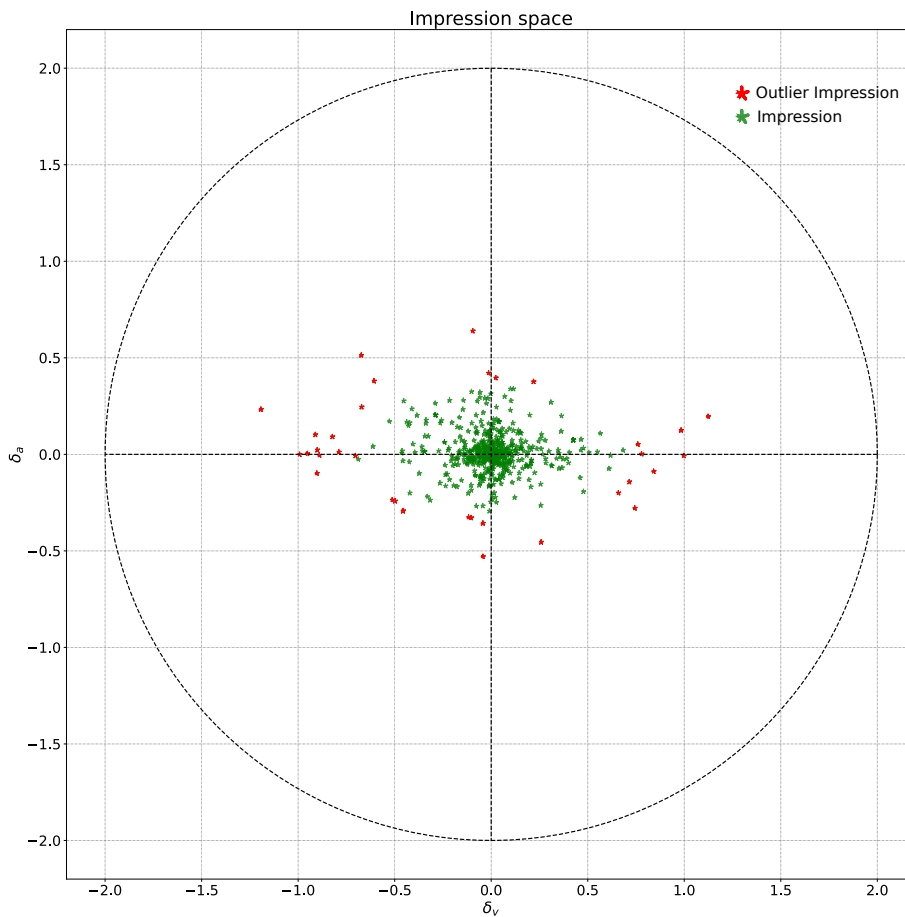


Fig. 5.6: Every emotion change $\vec{\delta}_E$ in the data set represented in the impression space.

From the emotional estimates obtained from each frame of every interaction with Kotaro, we have seen that the maximum valence shown was of 0.831, while the lowest valence was of -0.823 , while the average was -0.172 , with a standard deviation of 0.291. For arousal, the highest estimate

was of 0.815 and the lowest estimate was of -0.061 . The average arousal was of 0.176 with a standard deviation of 0.134.

Another metric that can be explored, for the video of a volunteer listening to a given GS utterance, is averaging the emotion estimates for each frame. This way, we can obtain an overall feeling of the emotion elicited by the interaction. Calculating such a metric for every video sample and averaging the average emotion, we obtained an average of all average emotion estimates $E_{avg_{avg}} = (-0.248, 0.161)$, with standard deviations of 0.293 for valence and 0.137 for arousal. The lowest and highest valence averages for each interaction were -0.693 and 0.831 , respectively. The lowest and highest arousal averages were -0.051 and 0.767 , respectively. Out of all 734 video samples, 130 video samples had a non-negative average valence and 695 had non-negative average arousal; and 37 video samples had both non-negative valence and arousal averages. Such results show that the majority of non-Yulean gibberish speech did generate moderately negative feelings on listeners, but still, few interactions had a positive average valence.

Considering participants that had more than a single exchange with Kotaro, it is possible to analyze how their emotional state changed along the overall interaction. Out of the 37 participants, 33 had multiple exchanges and 7 had interactions across multiple days. We used linear regression to detect the tendency of the evolution of the average valence of each interaction within the the same day (a participation session) and across multiple days, for the volunteers who participate multiple days (multiple sessions). Out of the 65 participation sessions, volunteers had their average valence decrease in 30 sessions, while in the remaining 35, the average valence of the interactions increased. For arousal, out of the 65 sessions, only in 28 could we see the valence increasing.

For volunteers who participated in multiple sessions across distinct days, the average valence decreased across different sessions for four volunteers, while it increased only for three of them. Arousal, on the other hand, increased only for two volunteers across multiple sessions, decreasing for the remaining five.

The results of such analysis can be seen in Figure 5.7, where the average valences of the different sessions are represented in different colors and the line resulting from the linear regression for each session has the same color as the points. The longest line represent the changes across multiple sessions. Some volunteers had their emotions improve, while other had their emotional state deteriorate while listening to the GS utterances. For the top left image, we can see a result were

the line fits the data very well, but for most volunteers, that is not the case, showing that continued interactions are not a good predictor of how well listeners will react to the different. Especially when looking at the bottom right graph of Figure 5.7, where we can see that in the first session there is a tendency of improving valence, but in the next session, there is a strong decrease in valence as the volunteer listened to the GS utterances.

For arousal, the results are similar, but since it has lower variance when compared to valence, linear fitting describes the evolution of the emotional state of participants during a session, as one can see in Figure 5.8. However, as it is possible to see in the bottom right figure, that is not the case for every volunteer. We have calculated the average mean squared error between the predicted and the actual impression for average arousal of every session, obtaining a value of 0.006, while the same metric for valence is of 0.023.

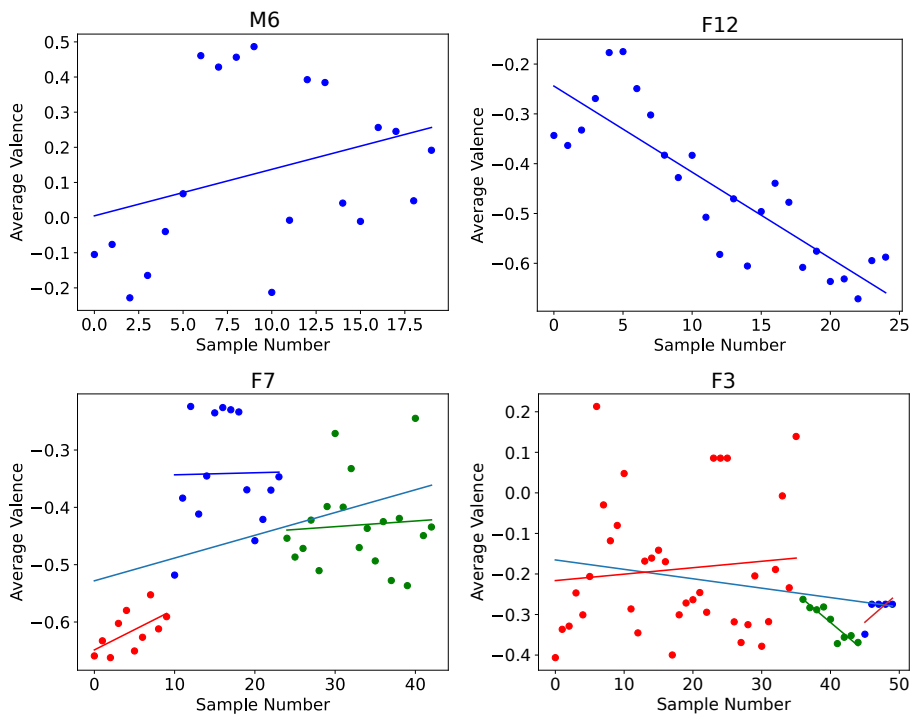


Fig. 5.7: Scatter plots of the average valence of each time volunteers M6, F3, F7, and F12 listened to a GS utterance and the results of the linear regression for each session and across multiple sessions. Points with the same color were obtained in the same session, and the line for that session shares the color with the points.

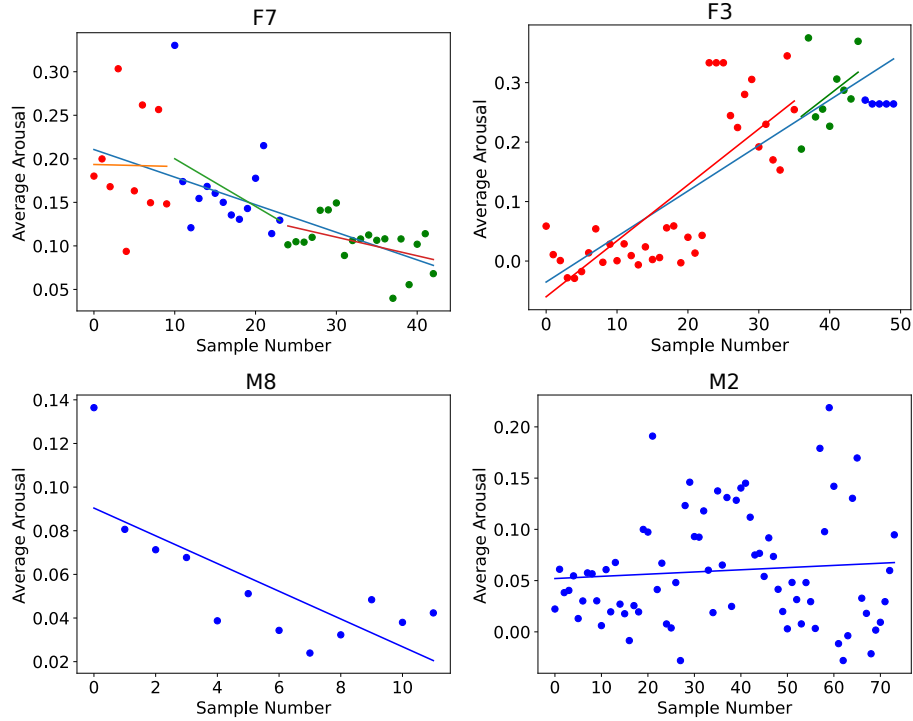


Fig. 5.8: Scatter plots of the average arousal of each time volunteers M2, M8, F3, and F7 listened to a GS utterance and the results of the linear regression for each session and across multiple sessions. Points with the same color were obtained in the same session, and the line for that session shares the color with the points.

Emotion State Change Estimate Error

In the context of the present work, the error of the estimate of the emotion change caused by gibberish speech $S(w, P)$ is defined as the norm of the difference between the actual emotional state change $\overrightarrow{\delta_{E_{S,A}}}$ and the predicted emotional state change $\overrightarrow{\delta_{E_{S,P}}}$, that is,

$$EE_S = \|\overrightarrow{\delta_{E_{S,A}}} - \overrightarrow{\delta_{E_{S,P}}}\| = \sqrt{(\delta_{v_{a,S}} - \delta_{v_{p,S}})^2 + (\delta_{a_{a,S}} - \delta_{a_{p,S}})^2}$$

Two different neural network architectures, VGG-16 and ResNet18, were used for estimating arousal and valence of volunteers from their facial expressions, respectively. However, since the aforementioned neural networks estimate human emotions from still image frames and the collected data consist of video samples, it was necessary to choose a metric that was capable of

representing the impact the gibberish speech had on the listener. This way, it is necessary to obtain the initial emotional state E_t of the volunteer and the emotional state E_{t+1} after listening to the utterance.

The chosen metric is then the difference between the emotion estimation from the initial (just before Kotaro starts speaking) and the last frames of each video sample. This metric is referred to as $\vec{\delta}_E = (\delta_v, \delta_a)$, where δ_v , δ_a is the change in the displayed valence and arousal. It is possible to see how a given utterance has affected research subjects in the valence–arousal space shown in Figure 5.4. If the valence improved, the vector is shown as blue; otherwise, as red.

The original research hypothesis during the development of the Talk to Kotaro platform was that prosodic choice is the most important factor in generating emotional responses in listeners; since gibberish has no meaning, it was expected that volunteers would respond according to prosodic features. Furthermore, it was expected that there would be a cross-cultural preference for certain prosodic parameters, similar to the Bouba–Kiki effect [15]. To test this hypothesis, it is necessary to compute the correlation between the prosody parameters and δ_v and δ_a . This analysis was performed pairwise using Stuart–Kendall’s τ_C correlation coefficient for each participant, all male volunteers, all female volunteers, all Japanese nationals, and all Brazilian nationals; their correlation matrices are shown in Figure 5.9. The correlation coefficient was also calculated for other demographics, but for the sake of brevity, the matrices are not shown.

It is very clear from Figure 5.9 that there is no statistically relevant correlation between the acoustic prosody characteristics and the generated impression for all volunteers, except for a very weak correlation between pitch and valence. For only the male participants, only the female participants, only Japanese nationals, and all Brazilian nationals as separate groups, no statistically relevant correlation could be found.

However, it is necessary to investigate if there are significant differences on the impressions displayed by men and women and by Brazilian and Japanese volunteers. In order to verify if the variance of the samples are similar, we performed multivariate analysis of variance, MANOVA, on the obtained data, whose results are shown in Tables 5.2 and 5.3. The results of column $\text{Pr} > F$ suggest that there is no statistically relevant difference between the reactions displayed by male and female volunteers. However, the reactions displayed by Japanese and Brazilian participants are statistically distinct.

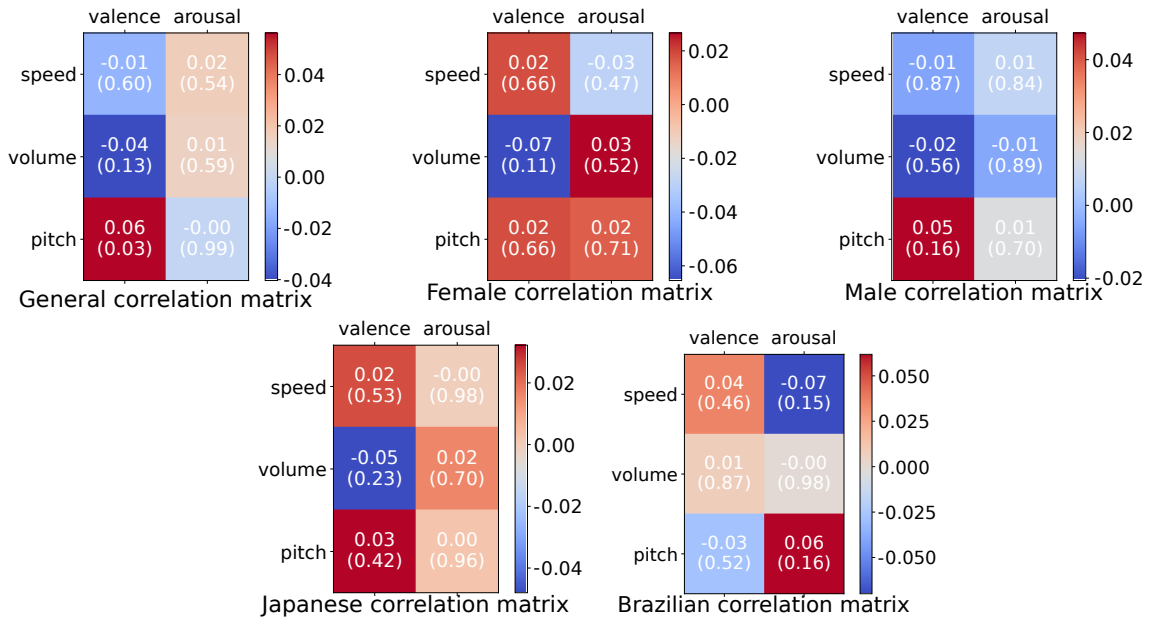


Fig. 5.9: Pairwise Stuart-Kendall's correlation coefficient matrices, where the top number of a cell indicates the coefficient and the number in parentheses indicates the related p -value.

Table 5.2: Results of the MANOVA for the data volunteered by male and female participants.

Group	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9950	2.0000	628.0000	1.5839	0.2060
Pillai's trace	0.0050	2.0000	628.0000	1.5839	0.2060
Hotelling-Lawley trace	0.0050	2.0000	628.0000	1.5839	0.2060
Roy's greatest root	0.0050	2.0000	628.0000	1.5839	0.2060

Table 5.3: Results of the MANOVA for the data volunteered by Japanese and Brazilian participants.

Group	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9716	4.0000	1120.0000	4.0687	0.0028
Pillai's trace	0.0285	4.0000	1122.0000	4.0534	0.0029
Hotelling-Lawley trace	0.0292	4.0000	670.9614	4.0889	0.0028
Roy's greatest root	0.0274	2.0000	561.0000	7.6855	0.0005

Thus, the original research hypothesis does not hold, i.e., there is no common baseline preference for particular prosodic patterns across cultures, across cultural groups, across genders, and

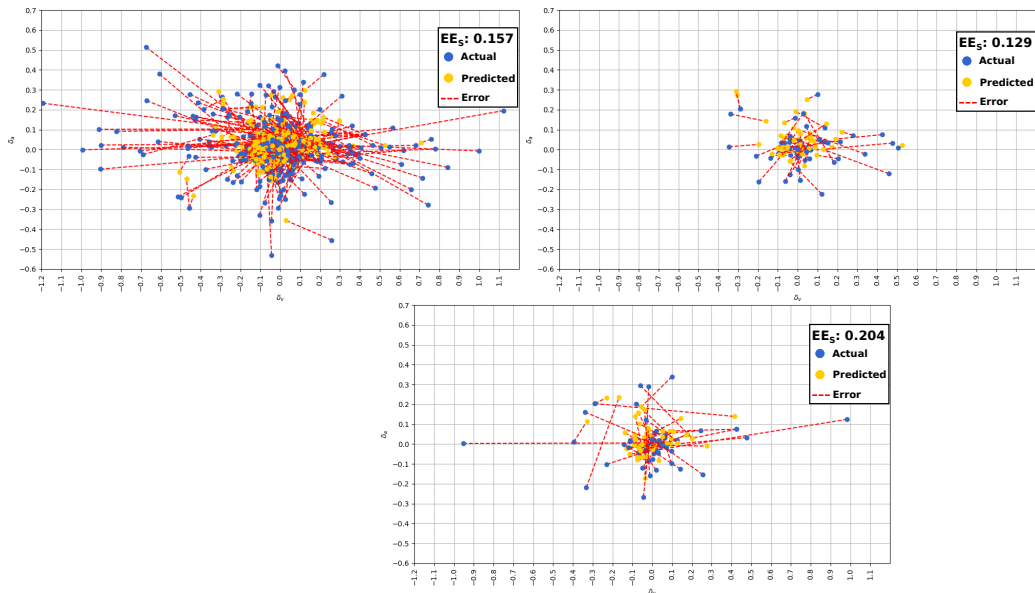
across age groups. Correlation between the acoustic prosody parameters and emotion change was investigated also for other groups, but since no other statistically relevant correlation was found, only the previously mentioned groups are displayed for the sake of brevity.

In order to strengthen the calculated correlations, assuming that some combination of the data points obtained through the “Talk to Kotaro” experiment actually reflect the real-world distribution of the reaction of how humans in general, men, women, Japanese people and Brazilian people would react to different acoustic prosody parameters, we perform statistical bootstrapping as defined in Section 2.1.5. In order to perform the bootstrapping technique, we consider the pairs (r, m) , where r is either speed, pitch, or volume, and m is either the associated δ_v or δ_a . During the Monte Carlo resampling operation, r and m are joined. After all sub-data sets are obtained, we separate all r and m into sets R and M and calculate the Stuart–Kendall τ_C correlation between both sets. For the present bootstrapping correlation analysis, we created 10,000 sub-data sets and used the percentile method to obtain the 95% confidence interval, whose results are shown in Table 5.4.

In order to obtain the desired *GSIP* model, we first removed the outliers from the data set and trained the $MLP_{profile+prosody}$ model using the prosodic characteristics of Kotaro’s remaining utterances and the profile of participants. Using the Adam optimizer (learning a rate of 10^{-3} , no decaying rate) with mean square error as the loss function, the model was trained for 100 epochs with a batch size of 32. The loss function for the training was mean squared error. Since the data set is quite small, 10% of the data were used for validation and 10% for testing. Two copies of the model were trained, one for valence and the other for arousal prediction. Together, they achieved an average error (as defined in Section 5.3.2) of 0.157 for the training data, 0.129 for the validation data, and 0.204 for the test data. The benchmarking results can be seen in Figure 5.10.

Table 5.4: Stuart–Kendall’s τ_C correlation 95% confidence interval obtained through bootstrapping.

Group	Prosodic Parameter	Valence	Arousal
General	Speed	[−0.065, 0.038]	[−0.036, 0.070]
	Volume	[−0.095, 0.014]	[−0.040, 0.068]
	Pitch	[0.0045, 0.110]	[−0.054, 0.052]
Male	Speed	[−0.076, 0.063]	[−0.061, 0.077]
	Volume	[−0.091, 0.050]	[−0.074, 0.064]
	Pitch	[−0.020, 0.114]	[−0.048, 0.076]
Female	Speed	[−0.062, 0.098]	[−0.119, 0.061]
	Volume	[−0.156, 0.025]	[−0.060, 0.114]
	Pitch	[−0.063, 0.101]	[−0.076, 0.106]
Brazilian	Speed	[−0.058, 0.127]	[−0.168, 0.029]
	Volume	[−0.087, 0.100]	[−0.100, 0.095]
	Pitch	[−0.117, 0.060]	[−0.018, 0.142]
Japanese	Speed	[−0.053, 0.103]	[−0.090, 0.086]
	Volume	[−0.136, 0.041]	[−0.067, 0.100]
	Pitch	[−0.051, 0.113]	[−0.083, 0.085]

Fig. 5.10: Comparison between the actual impression and the impression predicted by $MLP_{profile+prosody}$ for (top left) training data, (top right) validation data, and (bottom) test data.

Regarding the training of model GRU_{phones} , it is performed in Section 5.3.4, since it is also used for investigating the positioning of the phones in the phone embedding space.

5.3.3 Analysis of the recorded speech supports the findings of the video analysis

A total of 823 audio samples were recorded in the experiment, but many were unusable (were completely silent, contained very loud background noise *etc*), leaving us with 517 voice recordings. These voice recordings were analyzed using the LSTM-based neural network described in Subsection 5.1.2. The results are summarized in Table 5.5. It can be seen that the most frequent emotions of the recorded voice were disgust and anger, i.e. negative valence and low arousal values, and negative valence and high arousal values, respectively. These results are consistent with those obtained from the analysis of the volunteers' facial expressions, as shown in Figure 5.4. Happy, calm, and surprised initial states were rare but present in the interactions.

Table 5.5: Results of the sentiment analysis of volunteer's speeches.

Emotion Label	Number of Samples
Disgust	118
Angry	113
Happy	78
Surprised	54
Fearful	46
Sad	43
Calm	38
Neutral	27

Unfortunately, it was not possible to improve the accuracy of emotion estimation from the original video frames, but since the negative emotion estimates from the audio matched negative valence values and the positive ones matched positive valence values, it helped to validate, albeit qualitatively, the results of emotion estimation from facial expressions.

5.3.4 Phone Embedding Analysis

To investigate the contribution of each phone to the estimated impression across subjects, the GRU_{phones} neural network introduced in Section 5.1 was trained using the Adamax optimizer and

mean square error as the loss function with a batch size of 32 for 100 epochs. Two copies of the model were trained, one for valence and other for arousal, to estimate the change in arousal and valence caused by a given string of phones. The output dimension of the embedding layer was chosen after trials with many different values; the best results were obtained with an output dimension of 64. Thus, the resulting embedding matrices for valence and arousal are of 64×71 dimension. However, since the each phone has a position in a high-dimension hyperspace, it is not possible to visualize their proximity graphically.

To facilitate the analysis of the contribution of each phone, we used the k-means clustering method in order to group the phones accordingly to their proximity for both embedding matrices. In order to select the best number of clusters, we used the silhouette score analysis method [173] in order to determine the optimal number of clusters.

The optimal number of clusters was 35 for valence and 26 for the arousal embedding matrices. The number of phones in the largest cluster was 5, and most phones were grouped with distant phones or alone in their own cluster. From the clusters of the embedding spaces for valence and arousal, we were able to obtain Figures 5.11a,b, 5.12 and 5.13, which allows us to visualize which vowels and consonants phones belong to the same clusters, since such phones are colored with the same color. The red-colored phones were not selected by the gibberish speech generation algorithm, since the algorithm randomly picked phones.

However, since many phones were absent in Kotaro's gibberish speech, clustering might not be the best method to analyze which phones have a similar emotional impact on listeners, so we calculated the distance between the phones in the embedding space both for valence and arousal and for valence. Considering vowels and consonants separately, we obtained that for valence, the following phones that are close in their articulation loci are also the closest neighbors in the embedding space: [a]–[ɛ], [ʌ]–[ɛ], [ə]–[ɛ], [i]–[u], [p]–[m], [d]–[t], [k]–[g] and [ʃ]–[ʌ]. If we consider all phones together, we have all combinations of vowels and consonants, but none with close articulation in the human mouth.

We have performed the same analysis for the arousal embedding space, and fewer pairs have a close articulation locus in the human mouth were obtained than for valence. The following pairs were identified: [ɛ]–[e], [ə]–[ɛ], [o]–[u], [z]–[f], [ʒ]–[r] and [x]–[χ].

Such results show that while some phones are close in the learned embedding space for va-

lence, it is not possible to claim that similar phones, except for a few exceptions, are close in the embedding space, in the context of non-Yulean gibberish speech.

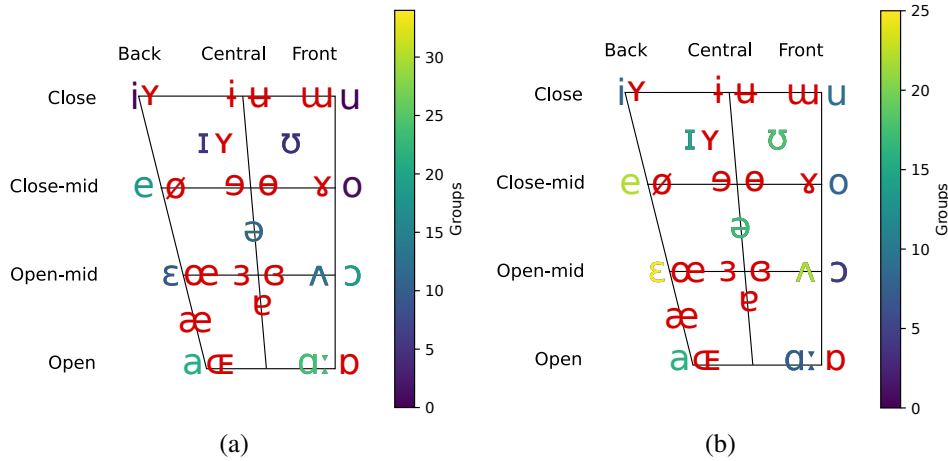


Fig. 5.11: Embedding values for vowels of the IPA for valence and arousal estimation. IPA symbols in red were absent in the generated utterances. Other symbols were colored according to the index of the cluster they belong to, as shown in the rightmost color bar; **(a)** Embedding values of vowel for valence change estimation; **(b)** embedding values of vowel for arousal change estimation.

	Bilabial	Labiodental	Dental	Alveolar	Post Alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			ʀ					ʀ		
Tap/Flap		ɸ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lat. Fricative				ɬ	ɮ						
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lat. Approximant				ɻ		ɻ	ʎ	ʎ			

Fig. 5.12: IPA Consonant table with embedding values for valence change estimation. IPA symbols in red were absent in the generated utterances. Other symbols were colored according to the index of the cluster they belong to, as shown in the rightmost color bar.

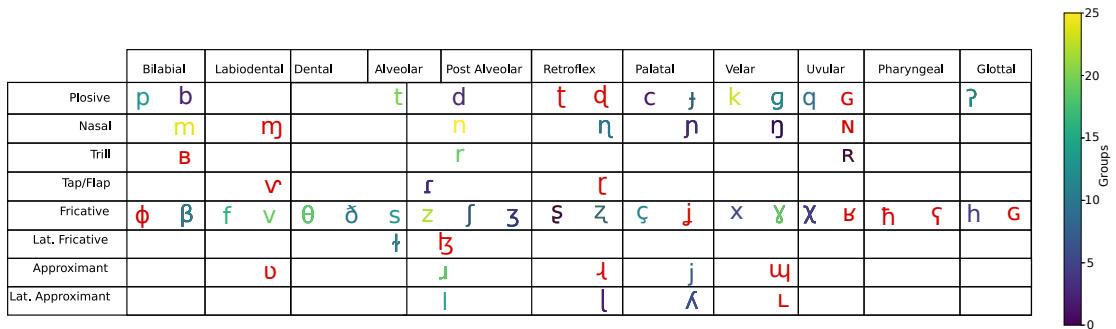


Fig. 5.13: IPA Consonant table with embedding values for arousal change estimation. IPA symbols in red were absent in the generated utterances. Other symbols were colored according to the index of the cluster they belong to, as shown in the rightmost color bar.

The proposed neural network was then able to estimate the emotional change caused just by the tokenized phone vector w of a given Gibberish speech $S(w, P)$, achieving a prediction error of 0.035 for training data and 0.241 for validation data, as one can see in Figure 5.14.

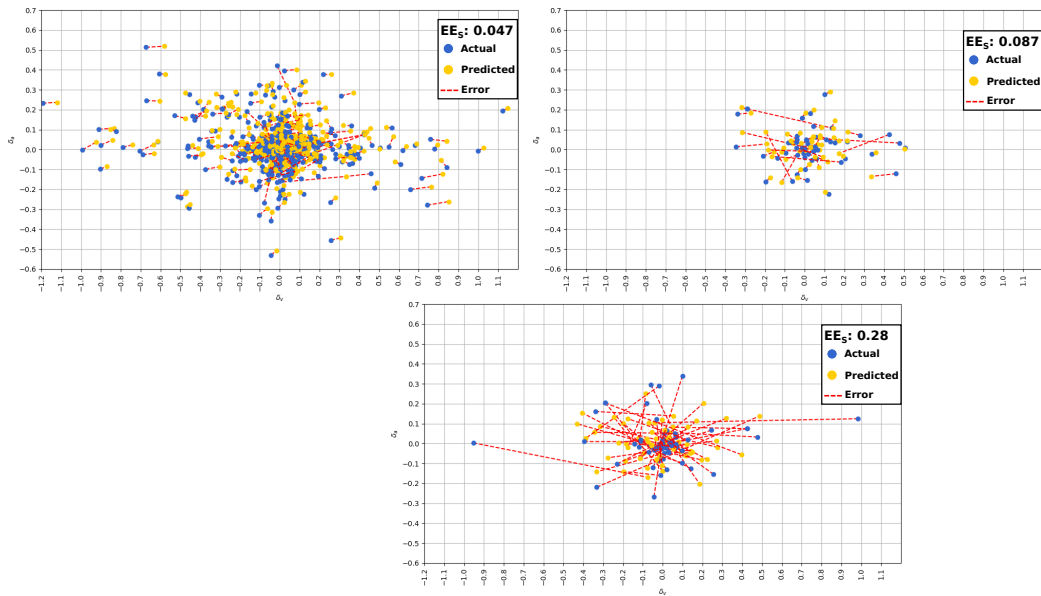


Fig. 5.14: Comparison between the actual impression and the impression predicted by GRU_{phones} for (top left) training data, (top right) validation data, and (bottom) test data.

5.3.5 Likert Scale Questionnaire Analysis

Out of a total of 37 research volunteers, only 22 (13 male and 9 female) answered the optional Likert scale questionnaire after participating in the experiment. With the results, it is possible to perform a *post hoc* analysis of the internal consistency of the questionnaire. Cronbach's alpha was chosen to measure the consistency of the questionnaire prompts; we obtained a Cronbach's alpha of 0.752, with a 95% confidence interval of [0.562, 0.881]. The internal consistency of the questionnaire is, thus, considered to be sufficient, and we can proceed with the analysis of the responses of the volunteers.

Given that prompts P_5 and P_8 were worded negatively, the responses must be inverted before any analysis is performed. Prompt P_3 , although seemingly negatively worded, does not change its meaning when inverted, i.e., if it had been worded as "Some randomly generated words are more pleasant than others", it would not have changed participants' responses, since some words being less pleasant than others already implies that some are more pleasant. The same is not true for P_5 and P_8 , which become "Different random words had an impact on your enjoyment" and "The turn-based conversation felt natural". To obtain the inverted responses IR from the actual responses AR , the following calculation must be made: $IR = MS - AR + 1$, where MS is the maximum score of the highest level of agreement; in this paper, it is 5.

To obtain the overall attitude toward a prompt, it is necessary to calculate the weighted average, where the value of a given item is multiplied by the number of respondents who chose that level of agreement, summed for each item, and divided by the total number of respondents in the questionnaire.

Results of the analysis performed on all prompts can be seen in the box and whisker plots shown in Figures 5.15 and 5.16, and the bar plots of each response by male and female volunteers can be seen in Figures 5.17 and 5.18. The overall attitude towards the Talk to Kotaro experiment was mostly neutral or slightly negative. Such results were expected after the emotion estimation analysis performed in Section 5.3.2, since most of the average emotion shown by participants during the experiment had negative valence and low but positive arousal.

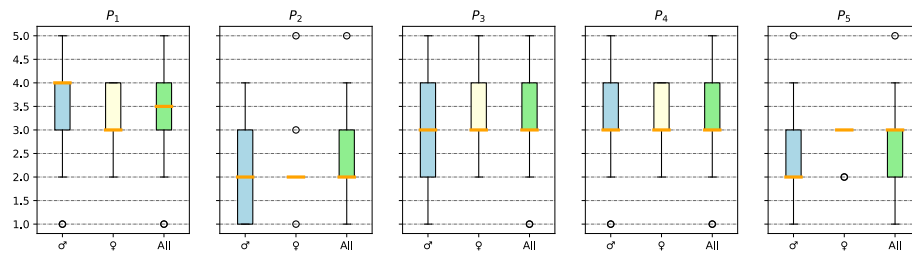


Fig. 5.15: Male (blue), female (yellow), and everyone's (green) responses to the optional Likert scale questionnaire's prompts 1 to 5. The median value of the responses is highlighted in orange, outliers are represented by small circles.

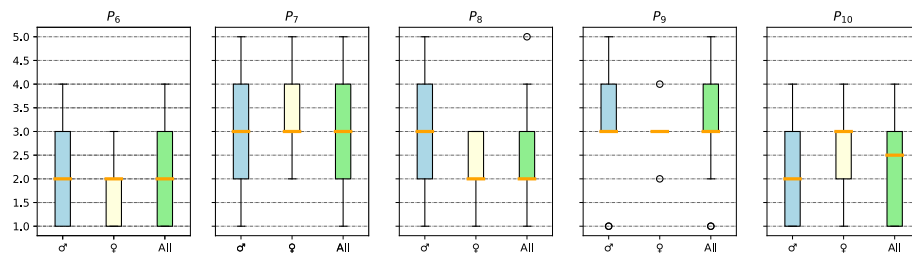


Fig. 5.16: Male (blue), female (yellow), and everyone's (green) responses to the optional Likert scale questionnaire's prompts 6 to 10. The median value of the responses is highlighted in orange, outliers are represented by small circles.

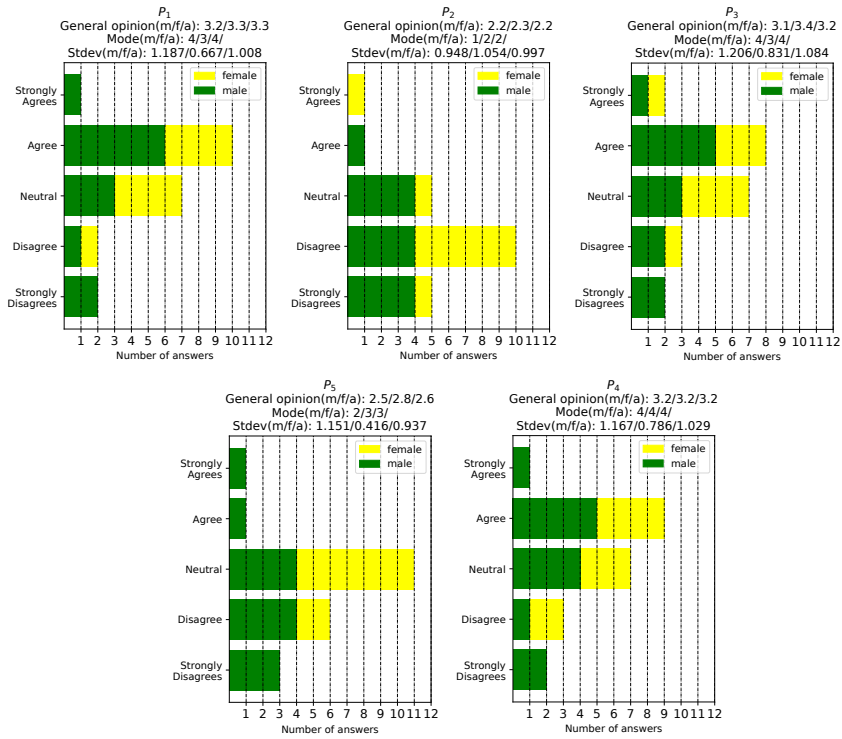


Fig. 5.17: Bar plots of the male and female responses to prompts 1 to 5 of the optional Likert scale questionnaire.

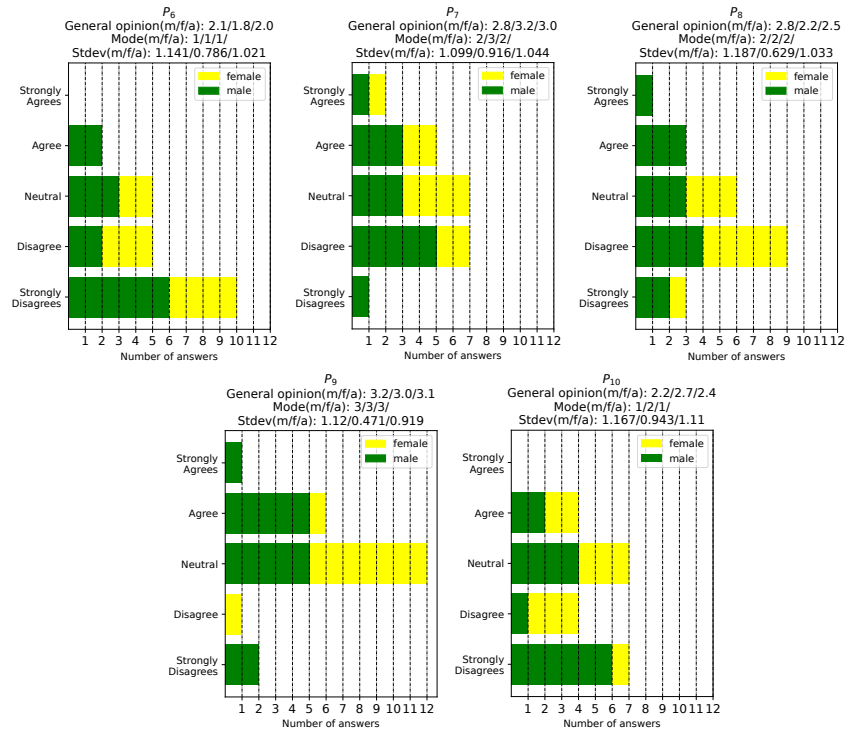


Fig. 5.18: Bar plots of the male and female responses to prompts 6 to 10 of the optional Likert scale questionnaire.

P_1 —Talking with the robot avatar was interesting The general opinion is that talking to the robot avatar was just slightly above neutral, but it must be noted that the most common answer for male respondents was actually that they agreed that talking to Kotaro was an interesting experience, while most women tied between finding it neutral or slightly interesting experience. There are many factors that could have contributed to such results, but in line with the opinions of participants in [56], talking to a gibberish-speaking robot does not lead to a very enjoyable conversation, even if it is more interesting than the nodding robot. However, it must be noted that the attitude towards the experience of talking to Kotaro was worse than the attitude towards talking with the Hanamogera-speaking NAO robot in [56]. Such result suggests that GS that has phone distribution tends to perform better than GS that does not. We are, however, cognizant of the fact that more embodied conversational agents tend to elicit higher engagement and better impression on research subjects, and thus, this result warrants further investigation.

P_2 —Variation of the Speech Characteristics Made Conversation More Natural Volunteers, both male and female, felt that the random prosody variations used by Kotaro for his gibberish speeches did not make the conversation feel natural. This result was somewhat expected, since the avatar could suddenly change its voice from a very high pitch to a very low pitch, sounding like a completely different entity. Such a result is supported by previous research works, such as [174], where pitch inflection is identified as a very important factor in voice recognition. Another point is that low volume and high speed may have affected the overall experience, since people usually do not suddenly change the speed or volume of their speech unless there is a context for doing so.

P_3 —Some randomly generated words are less pleasant than others The most frequent answer for participants was “3—neutral”, suggesting that research subjects could not see much difference on how distinct words generated by Algorithm 1 made them feel. This results suggests that non-Yulean gibberish speech words could not pick the interest of research subjects, again, in a similar fashion to the Hanamogera GS words in [56].

Moreover, since the algorithm also created some unusual combinations that were described by some participants to be “alien-like”, generated words might have caused estrangement on participants.

This result is also consistent with the data shown in Section 5.3.2, since most of the utterances left a neutral-to-negative impression on participants.

P_4 —Some speech characteristics, such as speed, loudness, or pitch influence more than others The analysis of this prompt was of particular interest since there was little to no correlation between speed, pitch, volume, and valence, and arousal. The question was somewhat divisive among the participants, since the most common response was “4—agree” (10 responses), although the seven neutral responses, “3—disagree”, and “2—strongly disagree”, skewed the overall attitude towards neutrality. The overall attitude agrees with the result of the emotion analysis from the video samples and with the lack of correlation between the acoustic prosody parameters and the impression of volunteers.

P_5 —Different random words didn’t have an impact on your enjoyment While the previous prompt analyzed the effect of prosody choice, this prompt analyzes the effect of phone choice.

It is very similar to prompt P_3 , but phrased differently to validate the results obtained. Since the prompt is negatively worded, in order to be comparable to the others, it is necessary to rephrase it as “Different random words had an impact on your enjoyment” and invert the responses.

The results were consistent for the overall attitude of all participants together and for female volunteers. However, for male volunteers, the overall impression worsened, since fewer male participants agreed with the random words. Such a result, even if unexpected, is more aligned with the emotion analysis from the video samples, but it shows that some volunteers might not be so sure of their opinion about the impact of phone choice in their impression.

P_6 —You felt that the robot was answering your speech accordingly This question, along with P_{10} , tests the perceived intelligence of Kotaro. The results indicate a highly negative perception, with the majority of male and female respondents strongly disagreeing with the prompt. The results suggest that the use of non-Yule-like distributions of phones and randomly changing prosody patterns leads to a poor opinion of the agent’s intelligence. Participants were likely aware of the random selection of phones and prosody patterns, which contributed to their negative perceptions.

P_7 —Longer phrases were more interesting The overall opinion that longer sentences are more interesting than shorter ones was rather neutral, but one can see that male participants had a worse attitude towards longer utterances, suggesting that men would prefer shorter gibberish utterances as a response.

P_8 —The turn-based conversation felt unnatural Another negatively worded prompt, P_8 , needs to be inverted to allow for a closer comparison with other prompts in the questionnaire. It then becomes “The turn-based conversation felt natural”, which tries to capture the effect that pressing a button to talk and having Kotaro answer might have had on the volunteers’ impressions. The general attitude is that the chosen turn-based conversation system felt unnatural. This was to be expected, since humans are very good at taking turns in conversation; the average silence between turns is within a range of 250 ms from the cross-language mean of 208 ms [175]. However, overall male impression was rather neutral, suggesting that such effect might be not as strong for male participants.

P_9 —Foreign sounding phones were more interesting The purpose of generating gibberish speech utterances using IPA symbols was to allow the ECA to use sounds from languages around the world, and this prompt was intended to measure the impact that foreign-sounding phonemes had on participants. The results indicate that attitudes toward foreign-sounding phonemes were mostly neutral, but an analysis of other responses shows that they were slightly more negative than positive.

P_{10} —The robot seemed to be intelligent Regarding the perceived intelligence of the embodied conversational agent, the results were mostly negative, in line with prompt P_6 , although not as much, since female respondents had “2—disagree” and “3—neutral” as the most frequent responses, while male responses were mostly “1—strongly disagree”. Again, subjects were able to perceive that the ECA randomly generated their responses. This prompt was phrased differently than P_6 to measure how the robot’s humanoid form affected perceptions of intelligence, since being intelligent and responding accordingly capture two different aspects of the ECA’s capabilities. While it did not improve the overall opinion of its intelligence, more responses were neutral, or even in agreement that the ECA was intelligent.

5.4 Evaluation of the GSIP

With both $MLP_{profile+prosody}$ and GRU_{phones} pre-trained, we further trained the combined models, using the standard gradient descent method (learning rate of 0.01 and no momentum) for 100 epochs with a batch size of size 32. It achieved an an average error of 0.141 for training data, 0.139 for validation data, and 0.19 for test data, as one can see in Figure 5.19.

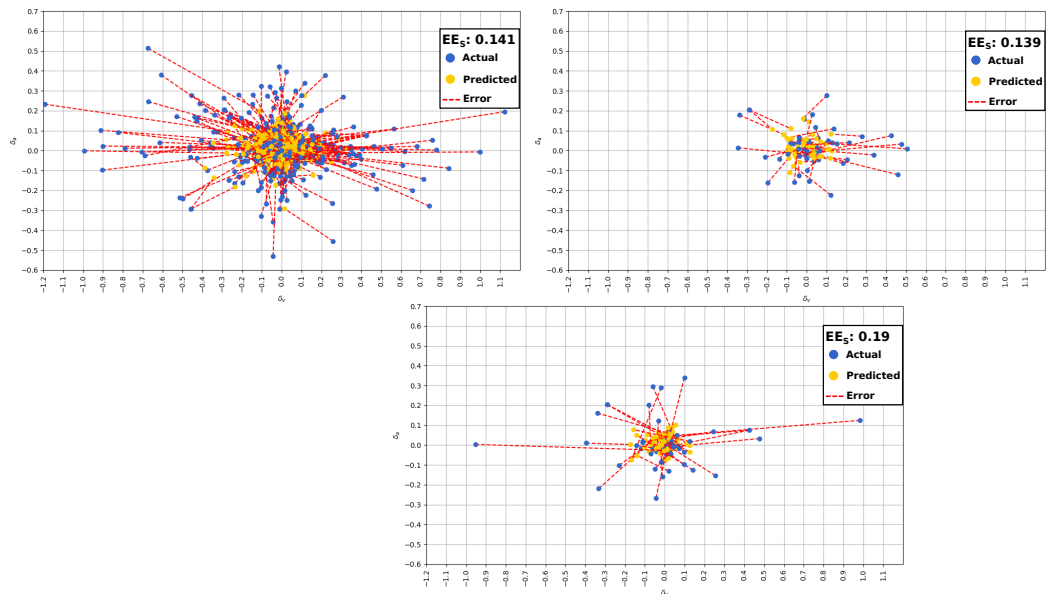


Fig. 5.19: Comparison between the actual impression and the impression predicted by *GSIP* for (top left) training data, (top right) validation data, and (bottom) test data.

5.5 Discussion

The results of the analysis performed on the audio and video recordings, together with the investigation on the location of each IPA phone in the learned embedding space and the results of the optional Likert Scale questionnaire, are individual pieces of a larger jigsaw puzzle that must be pieced together in order to allow us to see the bigger picture, enabling us to obtain further useful insights and to contextualize our previously shown results.

5.5.1 Effects of Kotaro’s Gibberish Speech on Listeners

The main takeaway when considering the results of the average emotion during the experiment, the impressions caused by the GS utterances, emotion classification of the audio samples, and the results of the Likert scale questionnaire is that GS that does not follow a traditional Yule-like phone distribution and has random acoustic prosody parameter selection does not have a good performance in a conversational setting with a screen-based conversational agent. Most utterances caused little to no impression, while the average emotion displayed while listening to Kotaro’s

utterances, $E_{avg_{avg}} = (-0.248, 0.161)$, could neither excite nor create positive feelings on listeners, on average. However, since the standard deviation for valence was quite high, $stdev_{valence} = 0.293$, we can see that there were still positive experiences, albeit few when compared with the neutral or slightly negative ones. Such results show that volunteers were mostly impatient and frustrated while listening to Kotaro's speech. There very few impression outliers, only 35 impressions, since most utterances caused small emotional changes.

Results of the analysis of what they told the agents also show that they showed little to no enthusiasm while talking to the agent, further showing that the overall experience was not particularly engaging. The neutral attitude toward prompt P_1 further sediments such a conclusion. Even though multiple participants have shown that they enjoyed through their answers, the majority still had a neutral or negative opinion of the experiment. Such results are in line with previous research results of work [56], where volunteers found the Hanamogera gibberish speech-speaking NAO robot more engaging than the nodding NAO robot, but volunteers still remarked that the conversations were still not so engaging. There was no acoustic prosody parameter variation in the GS utterances used by the NAO robot, and the overwhelmingly negative attitude towards prompt P_2 suggests that no variation of the prosody parameters performs better than completely random variations, as some volunteers also noted that drastic changes in pitch made them feel that they were speaking to a completely different entity, as voice pitch is a very important characteristic for identifying particular individuals just from their speech, as was shown in early voice identification works such as [174].

The main takeaway from the emotion analysis performed over the data provided by research subjects' suggestions and contrasting it with the results of previous research that focused on determining how research subjects felt regarding interacting with GS-speaking conversational agents [56, 100] is that while GS can provide positive interactions, its best use might not be in a conversational setting, since both in this work and in [56], volunteers complained about not understanding what the agent was saying and that they were not actually responding to their speech. Such results are unlike the ones shown in [100], where research subjects (children) played with a GS-speaking NAO robot a non-conversational setting, where the robot expressed its own emotions through GS. Since research subjects seemed to enjoy the experiment and to want to play again, GS in a expressive role (since the robot is using it to express itself) seems to perform better than

in a conversational setting, where more objective meaning is expected. However, another aspect to be taken into consideration is that in [100], research subjects were children, while in the present work and in [56], research subjects were mostly young adults who might be less accepting of such odd and “alien-like” interactions, since it requires a more imaginative and playful imagination, less focused in the actual communication and more in the experience itself.

Another reason that might explain the worse performance of the present GS generation technique is that the agent itself could not capture the interest of research subjects. The idea of making it mostly expressionless in a not-vibrant environment was to give more focus on the speech itself. Having an ECA on the screen was a deliberate choice to make the task actually resemble more the conversation with a robot or other types of ECA. Moreover, since higher embodiment levels tend to create higher engagement on users, we thought that having volunteers talk with a GS speaking voice without any representation would feel even less engaging, since research subjects could feel like they were talking to a non-entity. The researchers were, however, aware that the choice of the appearance of the ECA also matters in experiments, and the humanoid appearance of Kotaro might have created a mismatch in expected intelligence and the lack of coherence of the words said by the ECA, which tends to generate a bad impression on users, as discussed in [176] and exemplified by the lower perception of the robot in [177].

5.5.2 Effects of Prosody, Duration of Interaction, and Phone Choice

Previous analysis performed on video, audio, and Likert scale questionnaire answers can help us understand how research subjects felt towards each utterance and the experiment itself, but does nothing to elucidate why, which was one of the goals of the “Talk to Kotaro” experiment. In order to understand how acoustic prosody parameters affect the impression of volunteers, we have calculated pair-wise Stuart–Kendall’s τ_C correlation between each one of the investigated prosodic parameters and valence and arousal changes. Unlike what was previously thought by the researchers, no meaningful correlations could be obtained, with the exception of a very weak correlation of 0.06 between pitch and arousal for all participants, which had a p -value of 0.03.

By performing MANOVA analysis, it was possible to verify that Brazilian research subjects had distinct impression patterns compared to Japanese research subjects. Such analysis was not performed for other nationalities since they had too few participants (fewer than four), and thus,

it would be a meaningless comparison between an individual and a group of participants, for most nationalities. However, joining the fact that there were no favoured prosody patterns by all volunteers considered as a single group and that volunteers from different cultures had statistically distinct reactions to prosody parameters, we found no support for the original hypothesis from the “Talk to Kotaro Experiment” that, like the Kiki–Bouba effect [15], there would be a cross-cultural preference for certain prosody characteristics; quite the opposite.

However, it is necessary to further investigate if the high p-values are due to the small size of the data set or if there is really no statistically meaningful correlation. One way of performing such analysis is to use for all participants KDE (kernel density estimation) [178] to learn the distribution of the pairs of (p, r) , where p is one of the prosody parameters and r is the associated δ_v or δ_a value. With that, it is possible to create synthetic data whose distribution is very similar to the distribution obtained through the experiment and calculate the correlation between the synthetic set of p and r for different quantities of synthetic data points until meaningful p-values are obtained for all pairs of Stuart–Kendall’s correlation. Such an analysis is just a ballpark estimate, since it has a very strong assumption: the distribution of the real data obtained through the experiment actually represents (or represents closely enough) the actual distribution of how people react to different prosody parameters in the context of listening to non-Yulean GS.

We used a Gaussian kernel and chose a bandwidth of 0.1 to learn the distribution of our data in order to create synthetic data sets. With the distributions learned, we increased the size of the synthetic data sets until we consistently obtained meaningful, albeit still very weak, correlations between the synthetic pairs. We started obtaining mostly relevant correlations by 15,000 data points and always obtained statistically relevant correlations with 20,000 data points. Such a result shows that a much larger data set seems to be necessary in order to allow researchers to make stronger claims regarding the correlation between prosodic parameters and the impression of volunteers.

Linear regression was performed on the average emotion of each interaction volunteers had in their experiment sessions as a way of obtaining an overall tendency of how the emotion of participants evolved as they interacted with Kotaro. Both for valence and arousal, volunteers had positive or negative valence/arousal changes across the session, which are not explained by the number of interactions with Kotaro in a session, given that some volunteers that had multiple

sessions in different days had days where valence/arousal improved in one session and worsened on the next one, just to improve in the final session, as shown in the bottom right plot for F3 in Figure 5.7. Additionally, average valence values fluctuated a lot in a same session, very rarely showing any linear tendencies. Arousal, on the other hand, has shown better linear fit for most of the research subject, but not all. Moreover, even if the majority of research subjects showed decreasing arousal as they interacted with Kotaro, which is expected as the experience loses its novelty or as the participant gets tired, some research subjects showed increasing arousal, which is counter-intuitive. However, since users did not answer any personality tests or write any notes that could help elucidate the reason, it was not possible to understand why such patterns happened.

In order to analyze the position of individual IPA phones in the learned embedding space, we developed the GRU_{phones} neural network, which was able to learn to predict the impression of volunteers from Kotaro's GS utterances quite well for training and validation data, which shows good confidence on the 64×71 embedding spaces for predicating valence and arousal.

However, due to the stochastic nature of the algorithm, not every IPA phone was selected for the experiment. Moreover, both calculating the distances between the phones in the learned embedding hyperspaces and the clustering operations have shown no support for the idea that similarly sounding phones cause similar impressions, but many more data are necessary to lay stronger claims in that sense.

5.5.3 Performance of the GSIP System

In order to develop the gibberish speech impression prediction system, neural networks $MLP_{profile+prosody}$ (responsible for predicting impression just from the profile information of volunteers and the acoustic prosody characteristics of a GS utterance) and GRU_{phones} (responsible for predicting human impression from the tokenized IPA phones of a GS utterance) were pre-trained using the obtained data set after the outlier impressions were removed. GRU_{phones} achieved an outstanding performance for predicting training and validation data, but for test data, the results seemed lackluster and mostly random, which shows a lack of generalization capability of the model. For $MLP_{profile+prosody}$, the results were not as impressive for training and validation data, but it performed better than GRU_{phones} for test data, showing closer predictions for some of the test data.

By using the pre-trained neural networks, we trained the *GSIP* system, which consisted of the average of both previously mentioned neural networks, which achieved a better performance for test data when compared to previous two neural network, but it showed a tendency of making more “average” estimates, since most utterances generated small emotional changes.

The results were not satisfactory for test data, showing that even though the models could perform reasonably well for training and validation data, they could not properly learn how to generalize that knowledge for never-seen-before data.

第6章

GSIP Experiment: investigating the effects of Speech and Appearance

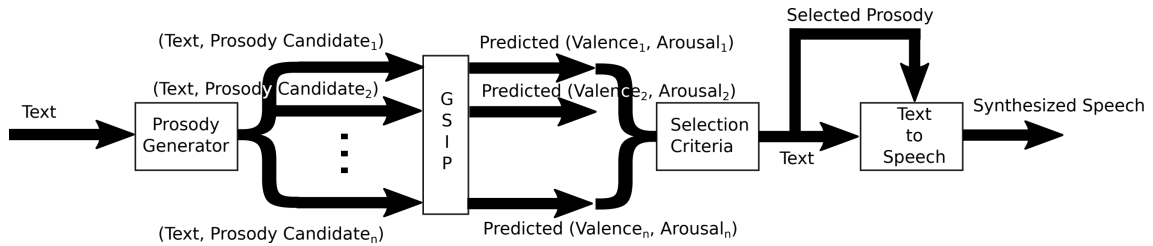


Fig. 6.1: Proposed architecture for the prosody Selection system.

6.1 Introduction

In order to make robotic speech more interesting to human listeners, the research titled “Image-based emotion detection system to improve human-robot communication” is currently being developed in Mizuuchi Lab. The first step of the aforementioned research consisted of the Talk to Kotaro experiment, where volunteers held conversation online with a robot avatar of the Kotaro robot. The website recorded what research subjects told the robot avatar and filmed their facial expressions while they listened to the semantic free utterances – speech composed from human phones but without meaning – Kotaro replied with. The experiment yielded enough data to allow the development of a module called GSIP – Gibberish Speech Impression Predictor. GSIP is capable of predicting the impression – immediate emotional response – of humans after listening to a semantic-free utterance spoken with a given prosody pattern. Its predictions are in terms of valence and arousal. Valence is the measure of how positive a given emotion is and it is defined over $\{v \in \mathbb{R} \mid -1 \leq v \leq 1\}$, -1 being the worst possible emotion and 1 being the most positive possible. Arousal measures how intense the displayed emotion is and its values are defined over $\{a \in \mathbb{R} \mid -1 \leq a \leq 1\}$, -1 being the most apathetic reaction and 1 being the most intense possible.

With the help of the GSIP system, proposed in Section 5.4, ECAs can better guess human impression to their utterances, allowing them to select adequate prosody parameters for pre-generated or randomly generated text of the semantic-free utterances (SFU). Batches of candidate prosodic patterns are randomly created and the [SFU, prosody] pairs are evaluated by the GSIP module – the pair which is predicted to generate the most positive impression (although other criteria can be chosen) is selected and spoken by the robot. The structure of the system is shown in Figure 6.1.

6.1.1 Research questions/hypotheses

The proposed system has been already developed but it is necessary to validate its performance experimentally and, thus, this Research Experiment Plan was devised in order to allow us to validate the hypotheses listed below:

- H_1 – speech generated by the proposed system is perceived as more human-like than speech with constant prosody or with randomly generated prosody;
- H_2 – speech generated by the proposed system generates more positive impression on volunteers than speech with constant or random prosody patterns;
- H_3 – the system could generate the desired impression in research subjects;
- H_4 – test subjects are more lenient with a non-humanoid looking avatar regarding semantic-free speech and eventual bad selection of prosody.
- H_5 – the system can be successfully used for physical robots;
- H_6 – the system can be successfully used for semantic speech;
- H_7 – novelty bias plays a heavy factor when users consider physical robots more engaging than virtual agents.

It is necessary to clarify that, in H_3 , generating the desired impression on volunteers is to obtain a valence-arousal impression with an mean absolute error (henceforth referred to as MAE) no larger than 0.15 (obtained with current data).

This research experiment plan, thus, explains the workflow of the experiments to be executed, what data will be recorded, how it will be recorded, stored and how research subjects will have their mental and physical integrity and anonymity protected.

6.2 Experiment Plan

6.2.1 General Overview

Since there are seven hypotheses to be experimentally tested for the developed system, the desired experiment is divided in three experimental phases P_1 , P_2 and P_3 , which were expected to last around 35 minutes each, totaling an 1h45 of experiment for each research subject. However there will be an initial explanation of the experiment where volunteers are free to ask any question. If they consent participating on the experiment, they will sign the consent form. As users may ask as many questions as necessary, it is hard to guess how long this preliminary phase might last but, if there are no questions, it is expected to last around 15 minutes. Between every experiment phase, participants will be given around 5 minutes to rest (more if they deem it necessary). This way, the total experiment time goes up to around 2h12, but in reality, most research subjects spoke with the conversation agents for farshorter periods of time and refused the first break, reducing the total average experiment duration to around 1h30.

In order to test hypotheses H_1 , H_2 , H_3 , H_4 and H_5 , the first experiment phase P_1 was designed. P_1 consists of having research subjects holding conversations with three different conversation agents A_1 , A_2 and A_3 , which will employ prosody selection systems $c_{1,gs}$, $c_{2,gs}$ and $c_{3,gs}$ for semantic free utterances. $c_{1,gs}$ always use the same prosody regardless of what the gibberish speech is; $c_{2,gs}$ is the system where prosody is selected using GSIP and, finally, $c_{3,gs}$ randomly selects prosody characteristics for semantic speech. Agent A_1 consists of the same avatar of the Kotaro robot used in the Talk to Kotaro Experiment; A_2 consists of a 2D avatar of the Plantroid Robot, which resembles an animal A_3 is the physical Plantroid robot. All conversation agents are shown in Table 6.1 The order in which volunteers talk to the agents will be randomly selected in order to avoid order bias; and the order on which the conversation agents will use the prosody selection systems will also be randomly selected for the same reason.

Every agent will hold a 10 exchange-long conversation for each one of the prosody selection systems, totaling 30 exchanges per agent and 90 in total for P_1 . An exchange is defined as a segment of a longer conversation, consisting of an initial saying by the research volunteer and a response from a conversation agent. Research subjects are free to say whatever they want to the conversation agent, which will reply using pre-generated semantic free utterances during the first

five exchanges. For the remaining five utterances, the first conversation agent will generate them using GSIP for the first prosody selection system it uses and they will be used for every other prosody selection system and conversation agent.

In order to test hypothesis H_6 , experiment P_2 will perform the very similar tests performed in P_1 , but the same conversation agents will use prosody selection systems $c_{1,ss}$, $c_{2,ss}$ and $c_{3,ss}$, where $c_{1,ss}$ applies the same prosody pattern for any semantic speech inputs, $c_{2,ss}$ uses GSIP to select the prosody characteristics of the speech and $c_{3,ss}$ randomly selects the prosody patterns. Volunteers will ask 3 given questions about the health of a plant and the conversation agents will answer with pre generated speeches. The remaining 7 exchanges are free; research subjects can talk about any topic with the conversation agents.

The final experiment phase P_3 consists of letting test subjects freely talk with conversation agents A_2 , A_3 and A_4 . A_4 is a holographic 3D model of the Plantroid robot shown using a Looking Glass holographic display. This experiment phase will test hypothesis H_7 , which postulates that the common perception that virtual agents are always less engaging than a physical embodied robot is because of novelty bias, which will be tested by introducing the novelty of interacting with a 3D hologram. If A_4 is considered to be as engaging or more engaging than A_3 , hypothesis H_7 will be validated. If the interactions with A_3 are found out to be more engaging than those with A_2 , but less than those with A_3 , it will show a strong support for novelty bias, but will validate the previous conception that physical embodiment of agents are more engaging than virtual agents. If it is found to be less engaging than A_2 , it will show that novelty plays little to no role into the interactions and a very strong support to the current conceptions of physical agents being more engaging. The prosody selection system for the responses of the different actors is $c_{1,ss}$, that is, same constant prosody.

After talking with every conversation agent, subjects will be asked to answer three questionnaires (Godspeed anthropomorphism and likeability) and will be asked if they want to rest, continue or give up the experiment. After all Phases, volunteers will be asked a final time for if they still consent to have their data used for the research. The experiment Phases will be described in more detail along with the questionnaires in the following sections. The flow of the experiment can be seen in Figure 6.2. All conversation agents are listed in Table 6.1 and every prosody selection system is show in Table 6.2.

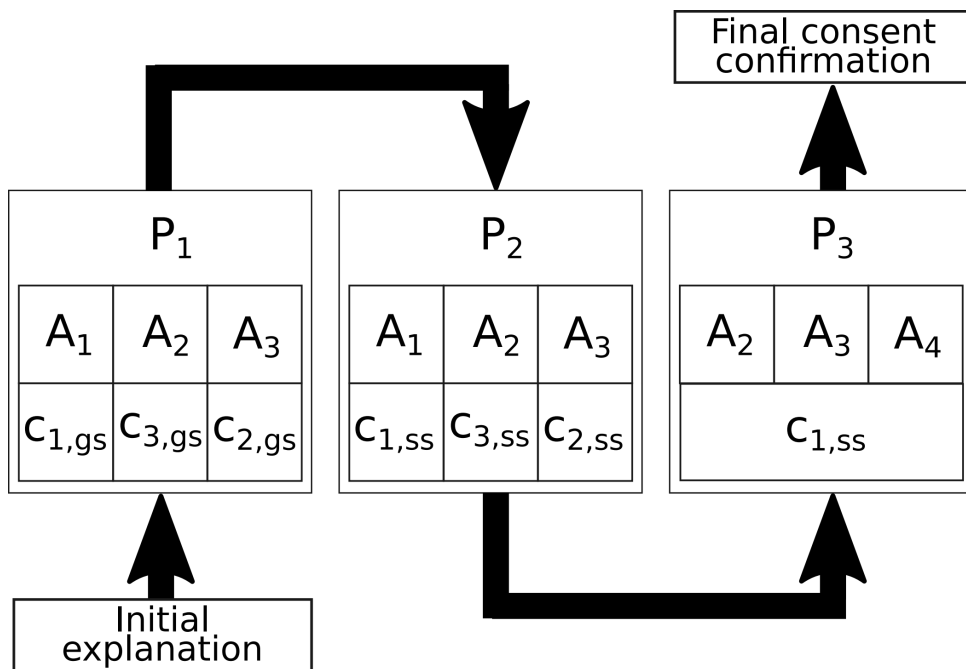


Fig. 6.2: Proposed experiment phases and its components.

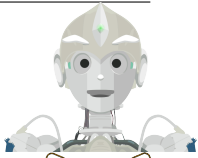
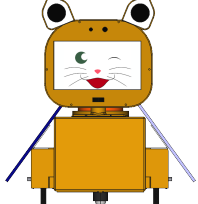
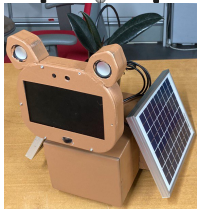

Name	Description	Image
A ₁	Avatar of the robot Kotaro.	
A ₂	Avatar of the robot Plantroid.	
A ₃	Physical Plantroid robot.	
A ₄	Holographic Plantroid robot.	

Table 6.1: Conversation agents, their descriptions and images.

Name	Description
$c_{1,gs}$	Constant prosody parameters for gibberish speech.
$c_{2,gs}$	Prosody parameters of gibberish speech selected using the GSIP system.
$c_{3,gs}$	Prosody patterns randomly selected for gibberish speech.
$c_{1,gs}$	Constant prosody parameters for semantic speech.
$c_{2,gs}$	Prosody parameters of semantic speech selected using the GSIP system.
$c_{3,gs}$	Prosody patterns randomly semantic for gibberish speech.

Table 6.2: Prosody parameters selection systems and their descriptions.

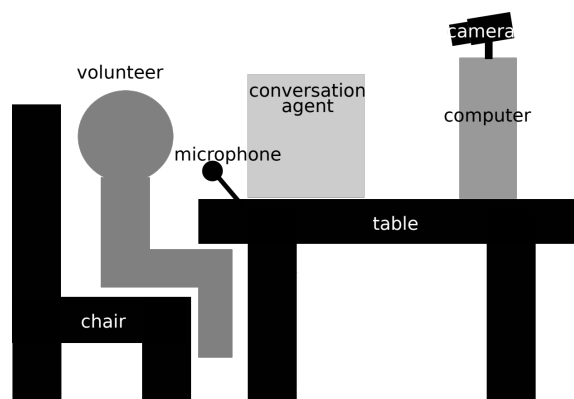


Fig. 6.3: Experiment Setup common for all phases.

6.2.2 Experiment Setup

The setup of the experiment, shown in Figure 6.3, consists of having a volunteer sitting down in front of a table where the conversation agent will be right in front of the eyes of the volunteer. Conversation Agents A_1 and A_2 will be shown in a LCD monitor. When Agent A_3 or A_4 are interacting with volunteers, the monitor will be removed and replaced by the physical Plantroid robot and by the Looking Glass holographic display. A computer, used for generating the prosody parameters and showing agents A_1 , A_2 and A_4 on their respective displays will also be above the table. An USB camera and a microphone will be placed in positions appropriate to capture the facial expression of users and their speeches while not obstructing their view of the conversation agents. Volunteers will be provided with a pen for answering questionnaires and signing the necessary forms. This configuration is common to every phase of the experiment.

6.2.3 Phase 1 (P_1)

Phase one consists of having the research volunteer holding ten exchange-long conversations with three distinct conversation agents A_1 (avatar of the Kotaro robot), A_2 (avatar of the Plantroid robot) and A_3 (the real Plantroid Robot), all of which will use three distinct prosody selection system, $c_{1,gs}$, $c_{2,gs}$ and $c_{3,gs}$ for their semantic free responses. In order to avoid any bias caused by having volunteers talk to the agent in the same order, the first agent with whom a volunteer will be selected among A_1 , A_2 and A_3 . For example, suppose that A_2 was selected as the first agent. A_2 will, then, randomly select a prosody generation system among $c_{1,gs}$, $c_{2,gs}$ and $c_{3,gs}$. Suppose $c_{1,gs}$ was chosen. The volunteer will have, then, ten exchanges with A_2 , which will reply for the first five times with pre-generated semantic-free utterances (showcased in Table 3), whose prosody will be always the same, accordingly to $c_{1,gs}$. The remaining five utterances will be randomly generated in this first interaction and reused for the subsequent interactions for the other prosody generation systems and for the other conversation agents. After the volunteer has talked with A_2 for ten exchanges, the subject has to fill a very brief Godspeed questionnaire, shown in Subsection 6.4.1. After that, the next prosody selection system will be randomly selected among the remaining two – the process is repeated for the next ten exchanges. Finally, the volunteer needs to talk to A_2 using the remaining prosody selection system. After filling the last Godspeed questionnaire, the volunteer will be asked to rank which prosody selection system he/she felt had the most human-like and most engaging speech. The same procedures will be repeated for the next conversation agent, which will be randomly selected among the remaining two. Finally, it will be repeated once again for the remaining agent. After interacting with every conversation agent, the volunteer will be asked to rank the most engaging conversation agent. An example of every step in of Phase 1 would be:

1. A_2 , $c_{1,gs}$ selected, volunteer interacts for 5 exchanges;
2. A_2 , $c_{1,gs}$ generates remaining 5 exchanges;
3. Volunteer answers Godspeed questionnaire about $c_{1,gs}$;
4. Selects A_2 , $c_{3,gs}$, volunteer interacts for 10 exchanges;;
5. Volunteer answers Godspeed questionnaire about $c_{3,gs}$;

6. Selects A_2 , $c_{1,gs}$, volunteer interacts for 10 exchanges;
7. Volunteer answers Godspeed questionnaire about $c_{1,gs}$;
8. Volunteer ranks most engaging prosody system;
9. A_1 , $c_{3,gs}$ selected, volunteer interacts for 10 exchanges;
10. Answer Godspeed questionnaire about $c_{3,gs}$
11. Selects A_1 , $c_{1,gs}$, volunteer interacts for 10 exchanges;
12. Answer Godspeed questionnaire about $c_{1,gs}$
13. A_1 , $c_{1,gs}$ selected, volunteer interacts for 10 exchanges;
14. Answer Godspeed questionnaire about $c_{1,gs}$
15. Volunteer ranks most engaging prosody system
16. A_3 , $c_{1,gs}$ selected, volunteer interacts for 10 exchanges;
17. Answer Godspeed questionnaire about $c_{1,gs}$
18. A_3 , $c_{3,gs}$ selected, volunteer interacts for 10 exchanges;
19. Answer Godspeed questionnaire about $c_{3,gs}$
20. A_3 , $c_{1,gs}$ selected, volunteer interacts for 10 exchanges;
21. Answer Godspeed questionnaire about $c_{1,gs}$
22. Volunteer ranks most engaging prosody system;
23. Volunteer ranks most engaging agent.

6.2.4 Phase 2 (P_2)

Phase 2 is quite similar to P1 in many aspects, however, conversation agents A_1 , A_2 and A_3 will reply using semantic speech, assigning prosody to it using systems $c_{1,ss}$, $c_{2,ss}$, and $c_{3,ss}$. Conversations with the (agent-prosody selection system) pairs consists of three exchanges where users will be request to ask the same predefined questions (show in Table 4), which will receive pre-generated responses and seven exchanges where they are free to say anything they want to the conversation agents, which will reply using a GPT-3-based chatbot. The chatbot used for every interaction is always the same. The remainder of the phase happens exactly like in phase 1 and an example of how a phase 2 could happen is given below:

1. A_1 , $c_{1,ss}$ selected, volunteer interacts for 3 exchanges;
2. volunteer freely interacts with A_1 , $c_{1,ss}$ for 7 additional exchanges;
3. Volunteer answers Godspeed questionnaire about $c_{1,ss}$;
4. Selects A_1 , $c_{3,ss}$, volunteer interacts for 3 exchanges;
5. volunteer freely interacts with A_1 , $c_{3,ss}$ for 7 additional exchanges;
6. Volunteer answers Godspeed questionnaire about $c_{3,ss}$;
7. Selects A_1 , $c_{2,ss}$, volunteer interacts for 3 exchanges;
8. Volunteer freely interacts with A_1 , $c_{2,ss}$ for 7 additional exchanges;
9. Volunteer answers Godspeed questionnaire about $c_{2,ss}$;
10. Volunteer ranks most engaging prosody system;
11. A_2 , $c_{2,ss}$ selected, volunteer interacts for 3 exchanges;
12. Volunteer freely interacts with A_2 , $c_{2,ss}$ for 7 additional exchanges;
13. Volunteer answers Godspeed questionnaire about $c_{2,ss}$;
14. Selects A_2 , $c_{3,ss}$, volunteer interacts for 3 exchanges;

15. Volunteer freely interacts with A_2 , $c_{3,ss}$ for 7 additional exchanges;
16. Volunteer answers Godspeed questionnaire about $c_{3,ss}$;
17. Selects A_2 , $c_{1,ss}$, volunteer interacts for 3 exchanges;
18. Volunteer freely interacts with A_2 , $c_{1,ss}$ for 7 additional exchanges;
19. Volunteer answers Godspeed questionnaire about $c_{1,ss}$;
20. Volunteer ranks most engaging prosody system;
21. A_3 , $c_{3,ss}$ selected, volunteer interacts for 3 exchanges;
22. Volunteer freely interacts with A_2 , $c_{3,ss}$ for 7 additional exchanges;
23. Volunteer answers Godspeed questionnaire about $c_{3,ss}$;
24. Selects A_3 , $c_{3,ss}$, volunteer interacts for 3 exchanges;
25. Volunteer freely interacts with A_3 , $c_{1,ss}$ for 7 additional exchanges;
26. Volunteer answers Godspeed questionnaire about $c_{1,ss}$;
27. Selects A_3 , $c_{2,ss}$, volunteer interacts for 3 exchanges;
28. Volunteer freely interacts with A_3 , $c_{2,ss}$ for 7 additional exchanges;
29. Volunteer answers Godspeed questionnaire about $c_{2,ss}$;
30. Volunteer ranks most engaging prosody system;
31. Volunteer ranks most engaging agent.

6.2.5 Phase 3 (P_3)

Phase P_3 is held in order to verify if novelty bias plays a heavy role on human impression regarding interacting with virtual agents and physical agents, that is an agent in a display and the physical robot. Current consensus is that embodiment is very important to make agents more engaging, however, there has been little consideration of the role of novelty bias in that conclusion,

that is, since most people have had very few interactions with robots, they consider interacting with a physical agent more interesting than interacting ones with screen-based virtual agents. To verify that, a holographic display will be used to show virtual agent A4, a 3D model of Plantroid robot. Since most people have not interacted with holographic agents, this introduces a novelty factor that might make research subjects feel more engaged while holding conversations with a virtual agent. In order to increase engagement, only semantic language exchanges will happen between the agents and research subjects. However, in order to not introduce prosody as an experience-changing factor, the prosody selection system $c_{1,ss}$ – constant prosody – will be used by all agents. Once again, to reduce order bias, the sequence on which research subjects will talk to the conversation agents will be determined at random and hold a five exchange long conversation. An example of how Phase 3 could happen is as follows:

- Selects A_2 , $c_{1,ss}$, volunteer interacts for five free exchanges;
- Selects A_4 , $c_{1,ss}$, volunteer interacts for five free exchanges;
- Selects A_3 , $c_{1,ss}$, volunteer interacts for five free exchanges;
- Volunteer answers Godspeed questionnaire about agents
- Volunteer ranks most engaging agent.

6.3 Collected Data

For the experiment, user information will be stored in many different forms. At the beginning of the experiment, a volunteer needs to fill two copies of the consent form, one which the research subject will take home and another which will be kept in Mizuuchi Lab, in a locker drawer. After that, it is necessary to fill a form about his/her personal information, specifically:

- Age;
- Gender;
- Country/Region of Origin;
- Mother Language;

- Other Languages you Speak;
- If you live or have lived abroad, write where;
- Years living abroad.

Such form will be filled in the computer program that was developed for the experiment. During the experiment, audio of what volunteers say to the conversation agents and of the replies of the conversation agents will be recorded by a microphone, while videos of the research subject's facial expressions will be recorded by a USB camera. After every interaction with a conversation agent using a given prosody selection system, volunteers need to fill an adapted Godspeed questionnaire; and after interacting with every prosody selection system, they need to rank which systems were the most engaging, human-like etc. After speaking to every conversation agent, research subjects need to answer a similar questionnaire, this time for ranking the conversation agents.

ID: _____

Please, rate your impression on how conversation agent ___ speech with system ___ sounds on the following scales:

Artificial ○○○○○ Natural

Unfriendly ○○○○○ Friendly

Unpleasant ○○○○○ Pleasant

Unintelligent ○○○○○ Intelligent

Apathetic ○○○○○ Responsive

ID: _____

Please, rate your impression on how conversation agent ___ speech with system ___ sounds on the following scales:

Artificial ○○○●○ Natural

Unfriendly ○○○●○ Friendly

Unpleasant ○○●○○ Pleasant

Unintelligent ○○○●○ Intelligent

Apathetic ●○○○○ Responsive

Fig. 6.4: Godspeed scale questionnaire for evaluating the impression caused by the speech characteristics of a prosody selection system.

6.4 Questionnaires

This appendix showcases the Godspeed questionnaires used for evaluating the different prosody selection systems and the conversation agents and the phrasing used in the form for ranking the same aforementioned prosody selection systems and agents.

6.4.1 Adapted Godspeed Questionnaire for Prosody Selection Systems

Research volunteers will receive a small paper card which looks like the one present in Figure 6.4 where they can evaluate their perception of the speech generated by a prosody selection system in the following scales: artificial-natural, unfriendly-friendly, unpleasant-pleasant, unintelligent-intelligent, apathetic-responsive. Research subjects will be handed out an example card on how to fill their card at the beginning of the experiment (also shown in Figure 6.4), which they will be able to consult at any moment. The fields "ID", "agent" and "system" will be filled by the researcher supervising the experiment after the volunteer has finished rating the prosody selection system, since the volunteers cannot know they will be talking with the same prosody selection system in different conversation agents (even though they might figure that out by themselves, they should not receive any confirmation in that regard, at least until the experiment is over).

The figure shows two identical-looking cards side-by-side. Each card has a rounded rectangular border and contains the following text and scales:

ID: _____
Please, rate your impression of agent ___ in the following scales:

Unfriendly ○○○○○○ Friendly
Unpleasant ○○○○○○ Pleasant
Unintelligent ○○○○○○ Intelligent
Apathetic ○○○○○○ Responsive
Kind ○○○○○○ Unkind

The right card shows an example of how to fill the scales. The filled circles are blue:

Unfriendly ○○○○○● Friendly
Unpleasant ○○○●○○ Pleasant
Unintelligent ○●○○○○ Intelligent
Apathetic ○○○○○● Responsive
Kind ○○○●○○ Unkind

Fig. 6.5: Left. Card for ranking the enjoyment of talking to each prosody selection system. Right. Example of how to fill such card.

6.4.2 Godspeed Questionnaire for Agents

Research volunteers will receive a small paper card which looks like the one present in Figure 6.5 where they can evaluate their perception of the conversation agent itself across distinct prosody selection systems in the following scales: unfriendly-friendly, unpleasant-pleasant, unintelligent-intelligent, apathetic-responsive, kind-unkind. Research subjects will be handed out an example card on how to fill their card at the beginning of the experiment (also shown in Figure 6.5), which they will be able to consult at any moment. The fields "ID" and "agent" are filled by the researcher supervising the experiment after the volunteer has finished rating the interaction with the agent, just to be consistent with how the system worked for rating the prosody selecting system.

6.4.3 Communication Systems Ranking

Participants will be asked to rank the performance of the prosody generation systems with which they just finished interacting with using a card shown in Figure 6.6. They will refer to the system as 1, 2 or 3, accordingly to the order on which they were used in the interactions with the current conversation agent. The researcher supervising the experiment will then attribute the correct names of the prosody selection systems after the volunteer hands out the card. While ranking the systems regarding how enjoyable the interactions were, subjects can use the comparison symbols $<$, $>$ and $=$, to denote that the system on the left performed worse than the system on the right, that

The figure shows two rounded rectangular cards side-by-side. Both cards have an 'ID: _____' field at the top. Below this, both cards contain the text: 'Please, rank the prosody selection systems in terms of how enjoyable were the generated speeches.' The left card has five blank lines for ranking: . Below these lines is the text 'system symbol system symbol system'. The right card has the same five lines, but they are filled with the example ranking: 3 > 1 > 2. Below these lines is the text 'system symbol system symbol system'.

Fig. 6.6: Left. Card for ranking the enjoyment of talking to each prosody generation system. Right. Example of how to fill such card.

the system on the left performed better than the system on the right and that the systems on the left and right had comparable performance, respectively. Volunteers will also receive an example card, shown in the right side of Figure 6.6 and explanation on how to rank the systems.

6.4.4 Agents Ranking

Participants will be asked to rank the performance of the conversation agents with which they just finished interacting with using a card shown in Figure 6.7. They will refer to the agents as 1, 2 or 3, accordingly to the order on which they were used in the interactions. The researcher supervising the experiment will then attribute the correct names of the agents after the volunteer hands out the card. While ranking the agents regarding how enjoyable the interactions were, subjects can use the comparison symbols $<$, $>$ and $=$, to denote that the agent on the left performed worse than the system on the right, that the agent on the left performed better than the system on the right and that the agents on the left and right had comparable performance, respectively. Volunteers will also receive an example card, shown in the right side of Figure 6.7 and explanation on how to rank agents.

The figure shows two identical cards side-by-side. Each card has a rounded rectangular border and contains the following text:

ID: _____
 Please, rank the conversation agents in terms of how much you enjoyed the conversation.

Below the text, there are three pairs of horizontal lines representing agents, with a 'symbol' between them. In the left card, the lines are empty. In the right card, the lines are filled with the following ranking: 1 > symbol > 3 = symbol > 2.

Fig. 6.7: Left. Card for ranking the enjoyment of talking to each conversation agent. Right. Example of how to fill such card.

6.5 Selected Semantic-Free Utterances, and Questions and their respective responses

In the Talk to Kotaro Experiment, every Semantic-Free Utterance (SFU) was generated by a Gibberish Speech (GS) generating algorithm, which caused every research subject to listen and react to distinct utterances. This, while helpful for understanding the role of phone selection on human impression, made analyzing how changing the prosody parameters impacts the impression of participants difficult. This time, however, five fixed GS were generated beforehand and, thus, every participant will react to the same SFU, allowing the researchers better analyze the role of personal preferences and prosody changes in human impression.

The same principle applies for having users ask previously selected questions which will receive semantic speech responses with the same text, so it is possible to better study the effects of different prosody parameters and of personal preferences.

This appendix will, then, list the GS in Subsection 6.5.1 and the questions and their respective responses in Subsection 6.5.2

IPA phones	speed	volume	pitch	valence	arousal
baɪ'ə'rɪu:'leɪb	150	100	45	0.05	0.2395
m,'a haθɹi:'w'ɛ ^θ ɪ	150	100	45	0.1079	0.3779
h,'aɪt'ɑ:ku:	150	100	45	0.1256	0.2048
ə'ʃ b a'l 'ɪm l c g	150	100	45	-0.405	0.206
q d'o	150	100	45	-0.05	0.3974

Table 6.3: Pregenerated Gibberish speeches and their predicted valence/arousal values.

6.5.1 Previously generated Gibberish Speech

All Gibberish Speeches listed in this subsection were randomly generated and, for prosody selection system $c_{1,gs}$, were evaluated by GSIP and, thus, have an expected impression prediction associated, allowing the researchers to compare with the actual impression displayed during experiments. All gibberish text (in IPA), volume, speed, pitch, predicted valence and predicted arousal are shown in Table 6.3.

6.5.2 Previously generated questions and their answers

The decision of having research volunteers asking selected questions and receiving selected responses from conversation agents has the same rationale of pre-generating the first five gibberish speeches of the agents, as explained in Subsection 2.1.1. However, the decision of having three defined question instead of five is in order to not let volunteers feel that their freedom was restricted, since in the previous experiment phase participants could say anything they wished to the robot. All questions, listed in Table 6.4, are related to the plant which Plantroid takes care of, since it is a neutral conversation topic that also gives purpose of existing to the conversation agents in the eyes of volunteers. For the same reason, every answer is either neutral or positive, in order to not negatively influence the participants' perception of the generated speeches.

Table 6.4: Pregenerated questions and their answers

Question	Answer
What is the current temperature?	The temperature in this room is of 26 degrees Celsius.
How long until I need to water the plant?	I estimate that you will need to water the plant in two hours and fifteen minutes.
What is the salinity of the soil?	The current salinity of the soil is of 2 deci-Siemens per meter; within the safe range for your plant.

第7章

GSIP Experiment - Execution and Results

7.1 Introduction

This chapter is dedicated to present the data obtained through the experiment explained in Chapter 6, to analyze it and discuss the results.

7.2 Experiment setup Implementation

To test all 7 research Hypotheses $H_{1,2,3,\dots,7}$, the experimental setup described in Chapter 6 was implemented. During phases $P_{1,2}$, volunteers engaged in brief open-ended conversations with Kotaro and Plantroid Avatars and Plantroid robot. The Holographic display was present in the table during the experiment, but it was kept turned off during first two phases to avoid distracting the volunteers. This arrangement, however, forced participants to take a break between Phases P_2 and P_3 , even when they did not want to do so, since it was necessary to turn on the holographic display and initialize the holographic Plantroid control script.

The interaction screen, which was used to show Plantroid and Kotaro avatars was implemented using Python and Kivy, appearing on a laptop computer's screen to volunteers. The same was done for the adapted Godspeed Scale questionnaire and Ranking questionnaires, the volunteers had to interact with the computer in order to evaluate the speech styles and the performance of the agents themselves.

A picture of the actual experimental setup can be seen of Figure 7.1.

Volunteers had to keep changing the focus of their attention between the Notebook screen, the robot and, during P_3 only, the holographic display, which made tracking their faces efficiently a little difficult, more than one camera should have been used for the experiment, considering that participants had to shift their line of sight and, thus rotate their heads frequently.

For Phases $P_{2,3}$, all conversational agents employed the same GPT3-based chatbot (OpenAI's GPT-3 davinci model), whose responses were synthesized into speech using the same speech synthesizer, eSpeak, and thus the only major difference between the agents is the level of physical embodiment (for P_2 and P_3) and prosody selection system (for P_2 only).

After having a conversation with an agent, participants had to fill out the Adapted Godspeed Scale questionnaire (presented in Subsection 6.4.1) about the speech style (during Phases $P_{1,2}$), another one after having interacted all speech style for an agent (presented in Subsection 6.4.2),



Fig. 7.1: text

rank the speech systems using the ranking questionnaire (presented in Subsection 6.4.4) finally, volunteers had to rank them in order of preference using the ranking questionnaire described in Subsection 6.4.4. Every questionnaire was filled in the notebook computer in the same Python Kivy program.

All audio was recorded with a condenser microphone, and the speech-to-text task was performed using Google's speech-to-text API. Although the Plantroid robot has its own microphone, in order to ensure equal performance for all agents, the robot was remotely controlled by the laptop computer by a simple TCP socket server, as it has more powerful hardware, thus avoiding potential delays in processing the audio data. The flow of the experiment is shown in Figure 7.3

7.2.1 Experiment limitations

The present experiment focuses during $P_{1,2}$ on the prosody selection systems, but some subtle differences between the agents might have had an impact larger than anticipated, such as: Kotaro is an humanoid robot, while Plantroid is pet-like, the size difference between the real robot and the avatars and the 3D model displayed in the holographic display. When interrogated about

the experiment after its completion, some participants noted that they did not enjoy interacting with the holographic display agent exactly because of its small size when compared with the real robot or screen agent. The extent of such effects cannot be fully known, since there was not a sliding scale of anthropomorphism and size to try and calculate how it impacts the performance of conversational agents. The effects of size are not addressed in this thesis, but the effects of the higher degree of anthropomorphism are partially addressed.

Moreover, regarding ranking the performance of the agents in Phase P_3 it is possible that the volunteer, having already had interacted with the Plantroid Avatar and robot for almost an hour, had already a formed opinion about them, while they had to judge the holographic agent after interaction.

In addition, the robot had its own speakers in its ears, which were quieter than the laptop speaker. Also, some participants indicated that they didn't enjoy interacting with the holographic agent very much because of its small size, which may have caused some more negative reactions to it. Finally, the conversations were open-ended, meaning that subjects could say or ask anything to the GPT-3-based chatbot. This can be seen as a limitation of the experiment, as participants were able to switch topics between agents, and the researchers acknowledge that a particularly interesting conversation could lead participants to prefer one agent over the others. However, when the participants were interviewed after the experiment, no participant mentioned that they preferred one ECA over the other based on the topic of the conversation, as they were instructed that they were dealing with the same AI.

Finally, it is a first-time interaction study and, thus, the results might not extend to continued interactions, since familiarity also plays a heavy role on interpersonal relationships [80]. However, even if that is the case the results still are useful, since it helps understanding first-time interactions better and, for many cases, users might interact with a system only once.

Moreover, a bug happened in the Python Kivy program and, for 6 of the participants, Kotaro did not show up in Phase P_2 , but none of the participants had noticed and still ranked Kotaro as if they have had an interaction with it during the experimental Phase (except the last one); all 5 of them ranked it as the better performing agent. That error shows two limitations of using the ranking scale: pre-conceived bias about the intelligence of agents based on their appearance and that human memory is very faulty. The problem was fixed, but that interesting phenomenon

Table 7.1: Adapted Godspeed Scale questionnaire for the Prosody Generation Systems.

S_1	Artificial	①	②	③	④	⑤	Natural
S_2	Unfriendly	①	②	③	④	⑤	Friendly
S_3	Unpleasant	①	②	③	④	⑤	Pleasant
S_4	Unintelligent	①	②	③	④	⑤	Intelligent
S_5	Apathetic	①	②	③	④	⑤	Responsive

shows the limitations of using a questionnaire after the interaction instead of using a metric that can measure the enjoyment of participants during the experiment, such as the audio and video recordings.

7.2.2 Adapted Godspeed Scale Questionnaire

The Godspeed Questionnaire Series is one of the most widely used human-robot interaction questionnaires to measure how research subjects perceive various characteristics of the robot during their interaction. It measures five essential HRI concepts – anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Respondents choose a value between two characteristics on a scale, e.g., unfriendly - friendly; lower values mean that the person perceives the robot as more unfriendly, and higher values show that they think the robot is friendly. The number of answers depends on how precise the researchers want the questionnaire to be, but at the cost of making it more difficult for participants to choose an answer.

The internal consistency of the scales has already been demonstrated, but due to the time constraints of the experiment and the fact that none of the five available scales $S_n, n = 1, 2, \dots, 5$ captured all the desired aspects of the interaction, we decided to combine certain items from different Godspeed scales and thus generated two adapted Godspeed questionnaire, one for the Prosody generation systems (shown in Table 7.2) and another one for the ECA (shown in Table 7.2).

Since this is a modification of the scales traditionally used, it is necessary to perform *post-hoc* tests to verify its internal consistency. Cronbach’s alpha was used to measure the internal consistency of both questionnaires for each phase, each ECA and each prosody selection system. Results of said analysis are shown in Tables 7.3 and 7.4:

Table 7.2: Adapted Godspeed Scale questionnaire for the ECA.

S_1	Unfriendly	①	②	③	④	⑤	Friendly
S_2	Unpleasant	①	②	③	④	⑤	Pleasant
S_3	Unintelligent	①	②	③	④	⑤	Intelligent
S_4	Apathetic	①	②	③	④	⑤	Responsive
S_5	Unkind	①	②	③	④	⑤	Kind

Most of the obtained internal consistencies were either good or acceptable, with only a few cases where the internal consistency is questionable, which are highlighted in red in Table 7.3; and all of the instances involved either GSIP or Random selection of prosody, raising the question if the participants could appropriately judge their impression of non-constant prosody.

For Phase P_3 considering all agents, the obtained Cronbach's alpha was 0.812, with a 95% confidence interval of [0.738, 0.87]. For the Plantroid robot agent, the obtained alpha was 0.788, with a confidence interval between [0.629, 0.892]. For the Holographic agent, the calculated alpha was 0.893, with a confidence interval between [0.718, 0.918]. Finally, for the screen agent, the alpha obtained was 0.807, with a confidence interval of [0.661, 0.902]. Thus, we can conclude that the adapted questionnaire has sufficient internal consistency for this study, but that we need to question the results obtained for GSIP and random prosody selection methods.

7.2.3 Ranking questionnaire

Besides the impression of the robot's characteristics, we needed to measure which agent was preferred by the participants, so a ranking questionnaire was created where users could enter symbols $>$, $<$ and $=$ to indicate which agents they thought were better or equal to others. An example of a response would be: robot $>$ holographic = screen.

In this way, it was possible to see which agent was preferred by the participants and to calculate the correlation between the rankings, the level of embodiment, and the users' experience with Robots in order to study the novelty preference and the level of embodiment in the participants' preference.

Table 7.3: Adapted Godspeed Scale questionnaire for the prosody generation systems internal consistency test for Phases $P_{1,2}$ and all subsets of combinations of agents and methods. $C\alpha$ stands for Cronbach's α and CI for Confidence Interval, Robot stand for Plantroid robot, Screen for Plantroid Avatar, rand for Random prosody selection, and const for constant prosody.

Phase	Agent	Method	$C\alpha$	CI
All	All	All	0.828	[0.802, 0.852]
All	Kotaro	All	0.8504	[0.807, 0.887]
All	Screen	All	0.790	[0.734, 0.839]
All	Robot	All	0.836	[0.791, 0.874]
P_1	All	All	0.784	[0.738, 0.826]
P_2	All	All	0.748	[0.69, 0.798]
P_1	All	const	0.821	[0.751, 0.877]
P_1	All	GSIP	0.724	[0.615, 0.811]
P_1	All	rand	0.794	[0.712, 0.859]
P_2	All	const	0.812	[0.734, 0.874]
P_2	All	GSIP	0.696	[0.569, 0.796]
P_2	All	rand	0.688	[0.555, 0.79]
P_1	Kotaro	All	0.791	[0.709, 0.857]
P_1	Screen	All	0.754	[0.656, 0.831]
P_1	Robot	All	0.801	[0.723, 0.864]
P_2	Kotaro	All	0.791	[0.709, 0.857]
P_2	Screen	All	0.754	[0.656, 0.831]
P_2	Robot	All	0.801	[0.723, 0.864]
P_1	Kotaro	const	0.822	[0.689, 0.91]
P_1	Screen	const	0.806	[0.657, 0.903]
P_1	Robot	const	0.837	[0.712, 0.918]
P_1	Kotaro	GSIP	0.767	[0.588, 0.883]
P_1	Screen	GSIP	0.638	[0.359, 0.819]
P_1	Robot	GSIP	0.749	[0.557, 0.875]
P_1	Kotaro	rand	0.783	[0.617, 0.892]
P_1	Screen	rand	0.765	[0.584, 0.882]
P_1	Robot	rand	0.818	[0.678, 0.909]
P_2	Kotaro	const	0.822	[0.689, 0.91]
P_2	Screen	const	0.806	[0.657, 0.903]
P_2	Robot	const	0.837	[0.712, 0.918]
P_2	Kotaro	GSIP	0.767	[0.588, 0.883]
P_2	Screen	GSIP	0.638	[0.359, 0.819]
P_2	Robot	GSIP	0.749	[0.557, 0.875]
P_2	Kotaro	rand	0.783	[0.617, 0.892]
P_2	Screen	rand	0.765	[0.584, 0.882]
P_2	Robot	rand	0.818	[0.678, 0.909]

Table 7.4: Adapted Godspeed Scale questionnaire for agents internal consistency test for Phases $P_{1,2}$ and all subsets of combinations of agents and methods. $C\alpha$ stands for Cronbach’s α and CI for Confidence Interval, Robot stand for Plantroid robot and Screen for Plantroid Avatar.

Phase	Agent	$C\alpha$	CI
P_1	All	0.825	[0.755, 0.88]
P_2	All	0.741	[0.632, 0.826
P_1	kotaro	0.853	[0.74 , 0.926]
P_1	screen	0.808	[0.661, 0.904]
P_1	robot	0.827	[0.694, 0.913]
P_2	kotaro	0.770	[0.545, 0.903]
P_2	screen	0.705	[0.484, 0.85]
P_2	robot	0.758	[0.572, 0.879]

7.2.4 Emotion Analysis

In order to estimate the emotional state of the participants while interacting with the conversational agents, all interactions were filmed using a 4k camera. The captured frames of each video were analyzed using two deep neural networks, VGG-16 and ResNet-18 [163], trained on the AffectNet dataset [164], to estimate the emotional state from the participants’ facial expressions in the 2D valence-arousal emotion space [68]. However, such a network can only estimate the valence (how positive the displayed emotion is) and arousal (the intensity of the emotion) of the facial expression present in a single frame, resulting in an emotion time series, as shown by the blue line graphs in Figure 7.4. However, since the estimates are quite noisy, a moving window of 0.5 seconds is used to smooth the time series. Since it is difficult to capture the full emotional dynamics of the valence and arousal curves with a single value, we initially suggested using the integral of the valence and arousal curves, as it captures the overall emotional state over time, as shown in the red area of the graphs in Figure 7.4.

However, such approach is valid only if the length of the generated utterances does not significantly change, or else an utterance that is less impactful, but lasts longer, might be considered better/worse than a more impactful, but briefer utterance. The objective is not measuring how much emotion is stored in a video, like calculating the power of a signal. The goal is to understand what emotions the utterances are generating and, thus, calculating the average emotion displayed in the frames of a video and the variance might be a better approach, since it is a duration-independent

metric. However, such approach is also not ideal, since a very strong positive/negative estimation might skew the final average. Nonetheless, when different prosody systems that generates utterances with different duration are involved; and given that the content of the utterances themselves is not fixed, it is better to use the average of the emotion displayed in every frame, at least for Phases 1 and 2.

7.3 Profile of participants

The experiment was conducted at the Tokyo University of Agriculture and Technology (hereafter referred to as TUAT) between February 13th and March 3rd of 2023, with a total of 28 participants, of which 27 could have their data used. Of these 27 participants, 14 were female and 13 were male. All participants were recruited from TUAT students, with a mean age of 25.56 years (standard deviation: 4.05 years). The youngest participant was 19 years old and the oldest was 33 years old. Of the total 27 participants, 14 had no previous experience with robots (level 0), 5 had little experience (level 1), 3 had intermediate experience (level 2), and 5 had a lot of experience with robots (level 3). The nationality that had the most participants was Japanese, with a total of 5 volunteers, and Malaysian was second with a total of 4 volunteers. Most participants had some experience with pets, which is represented by a simple binary encoding, 0 implies that the volunteer never had pets and 1 implies they have had experience with pets. Their regions of origin, places of stay abroad and languages they are capable of speaking are listed in Table 7.5

7.4 GSIP experiment Phase 1 - Gibberish Speech

The first phase of the GSIP experiment was designed to test the following research hypotheses:

- H_1 – speech generated by the proposed system is perceived as more human-like than speech with constant prosody or with randomly generated prosody;
- H_2 – speech generated by the proposed system generates more positive impression on volunteers than speech with constant or random prosody patterns;
- H_3 – the system could generate the desired impression in research subjects;

Table 7.5: Profile of participants breakdown.

Region of Origin	Total	Language	Total	Where Abroad	Total
Brazil	1	Indonesian	4	USA	1
Hong Kong (China)	1	Malay	4	Netherlands	1
China	3	Portuguese	1	India	1
Malaysia	4	Thai	3	United Kingdom	1
Thailand	2	Chinese	5	New Zealand	1
Myanmar	1	Cebuano	1	Australia	1
Mongolia	2	Burmese	1	Japan	20
France	1	Mongolian	1	Experience with	
Ghana	2	Cantonese	2	Robots	Total
Tunisia	1	Japanese	22	0	14
Philippines	1	Akan	2	1	5
Indonesia	3	French	2	2	22
Japan	5	Ga	1	3	5
Gender	Total	German	1	Experience with	
Male	13	Ewe	1	Pets	Total
Female	14	Mandarin	1	0	6
Age groups	Total	Filipino (tagalog)	1	1	21
18<age<20	2	Arabic	1		
20<=age<30	20	Mongolian	1		
30<=age<40	5	English	27		

- H_4 – test subjects are more lenient with a non-humanoid looking avatar regarding semantic-free speech and eventual bad selection of prosody.
- H_5 – the system can be successfully used for physical robots;

If Hypothesis H_1 , and H_2 are valid, we expect that agents using the GSIP prosody selection system will have better evaluation in the adapted Godspeed Scale questionnaire than when using other systems; and will be ranked higher than other prosody selection systems in the ranking questionnaires. Regarding H_3 , we need to verify the error between the predicted impression and the impression that was actually generated. Since the impression is modeled, in the context of this work, as a vector that represents the immediate change in the emotional state of volunteers, the error can be calculated as the norm of the difference vector between the actual emotion change and the estimated emotion change; that is:

$$I_{error} = ||\overrightarrow{I_{actual}} - \overrightarrow{I_{GSIP}}||$$

For H_4 , we need that the overall evaluation for random prosody selection and for GSIP-based prosody selection of non-humanoid-looking agents are better than those of Kotaro robot with the same systems and that the difference of perceived intelligence reduces less for such agents than for Kotaro, when the system is changed from constant to the other two.

Finally, in order to H_5 be true, Plantroid robot should have similar or better perception of Plantroid avatar in the adapted Godspeed Scale questionnaire.

The analysis of the video recording can be used to estimate the impression caused by the agents during the interactions and, thus, is considered to be a more accurate representation of the enjoyment of the interactions, albeit it does not allow to gauge the perception of particular items captured by the questionnaire.

With that in mind, it is necessary to analyze the responses to the questionnaires and obtain the overall emotion caused by the utterances of the ECA.

7.5 GSIP experiment Phase 2 - Semantic Speech

The second phase of this experiment was designed mostly to investigate hypothesis H_6 . but also to study how differently research subjects perceive the ECAs when they switch from Gibberish

speech to semantic speech.

7.6 Godspeed scale questionnaires response analysis

In order to investigate the perception of participants about all prosody generation systems, research volunteers had to evaluate how friendly, pleasant, intelligent, responsive and kind did the speech generated using said systems. In order to gauge the overall perception, we obtain the median of the scales, considering all combinations that made sense, for example, investigating the perception for all phases and systems for a given agent makes no sense for the questionnaire used for measuring the perception of the systems; such analysis must be performed by the adapted Godspeed Scale questionnaire for the agents. The results for the systems are shown in Table 7.6 and the results for the ECA are shown in Table 7.7.

From the data shown in Table 7.6, it is possible to notice that the generated speech for all systems and agents is perceived as rather unnatural, since no combination of Agent and system had an overall attitude over 3; and only Kotaro and the Plantroid robot achieves so while using constant prosody. The same can be said for the perceptions of Friendliness and Pleasantness, all with a rather neutral score of 3. However, for intelligence, the generated gibberish speech is perceived as unintelligent, albeit as not completely so. However, all systems except the proposed GSIP, received a good evaluation of responsiveness (score of 4), showing that volunteers noticed the slightly longer delay between finishing speaking and the ECAs answering, a delay caused by GSIP taking between 2 and 5 seconds to attribute prosody to the IPA phone inputs, which had a neutral general perception of 3. It is possible to see that there is an improvement on overall perception of the generated speech no matter the agent or the prosody selection system between Phases P_1 and P_2 .

Now, analyzing Table 7.7, once can see the perception about the agents themselves more directly. Kotaro was considered the least friendly in the scenario where we consider all phases together; but has the same neutral perception of other agents during P_1 . However, the perception of Plantroid Avatar and Plantroid Robot improve during phase P_2 , which does not happen for Kotaro. That can be due to the fact that Plantroid is styled like a cute pet, while Kotaro is a machine-like humanoid. No agent was considered to be unpleasant; and we can see that all agents go from a neutral perception to good perception of pleasantness from between the phases. The

Table 7.6: Median and Variance of the responses to the Prosody Selection System Adapted God-speed Scale Questionnaire of phases $P_{1,2}$ and all possible combinations of agents and prosody selection systems

Phase	Agent	System	Artificial/Natural	(Un)Friendly	(Un)Pleasant	(Un)Intelligent	Apathetic/Responsive
All	All	All	3.0/1.195	3.0/1.025	3.0/1.19	3.0/1.758	4.0/1.213
P_1	All	All	2.0/1.053	3.0/1.08	3.0/1.195	2.0/1.182	4.0/1.526
P_2	All	All	3.0/1.132	4.0/0.655	4.0/0.706	4.0/0.804	4.0/0.619
P_1	All	const	2.0/1.116	3.0/1.045	3.0/1.223	2.0/1.268	4.0/1.612
P_1	All	GSIP	2.0/0.989	3.0/1.068	3.0/1.16	2.0/1.102	3.0/1.525
P_1	All	rand	2.0/1.056	3.0/1.145	3.0/1.232	2.0/1.2	4.0/1.357
P_2	All	const	3.0/0.891	4.0/0.678	4.0/0.625	4.0/0.751	4.0/0.54
P_2	All	GSIP	3.0/1.095	4.0/0.539	4.0/0.592	4.0/0.92	4.0/0.755
P_2	All	rand	3.0/1.293	4.0/0.716	4.0/0.828	4.0/0.695	4.0/0.54
P_1	kotaro	const	2.0/1.103	2.0/0.866	3.0/0.949	2.0/1.37	3.0/1.738
P_1	screen	const	2.0/0.974	3.0/0.954	3.0/1.545	2.0/1.175	4.0/1.645
P_1	robot	const	3.0/1.294	3.0/1.282	3.0/1.226	2.0/1.342	4.0/1.46
P_1	kotaro	GSIP	2.0/0.794	2.5/0.894	3.0/0.978	2.0/1.095	3.0/1.655
P_1	screen	GSIP	2.0/0.962	3.0/1.225	3.0/1.355	2.0/0.918	3.0/1.466
P_1	robot	GSIP	2.0/1.262	3.0/1.145	3.0/1.12	2.0/1.342	3.0/1.575
P_1	kotaro	rand	2.0/0.906	3.0/0.98	3.0/0.978	2.0/1.175	4.0/1.38
P_1	screen	rand	2.5/1.138	3.0/1.095	3.0/1.466	2.5/1.165	4.0/1.595
P_1	robot	rand	2.5/1.134	3.0/1.226	3.0/1.226	2.0/1.342	4.0/1.182
P_2	kotaro	const	2.0/1.103	2.0/0.866	3.0/0.949	2.0/1.37	3.0/1.738
P_2	screen	const	2.0/0.974	3.0/0.954	3.0/1.545	2.0/1.175	4.0/1.645
P_2	robot	const	3.0/1.294	3.0/1.282	3.0/1.226	2.0/1.342	4.0/1.46
P_2	kotaro	GSIP	2.0/0.794	2.5/0.894	3.0/0.978	2.0/1.095	3.0/1.655
P_2	screen	GSIP	2.0/0.962	3.0/1.225	3.0/1.355	2.0/0.918	3.0/1.466
P_2	robot	GSIP	2.0/1.262	3.0/1.145	3.0/1.12	2.0/1.342	3.0/1.575
P_2	kotaro	rand	2.0/0.906	3.0/0.98	3.0/0.978	2.0/1.175	4.0/1.38
P_2	screen	rand	2.5/1.138	3.0/1.095	3.0/1.466	2.5/1.165	4.0/1.595
P_2	robot	rand	2.5/1.134	3.0/1.226	3.0/1.226	2.0/1.342	4.0/1.182

Table 7.7: Median and Variance of the responses to the Adapted Godspeed Scale Questionnaire for Agents of phases $P_{1,2}$ and all possible combinations of agents and prosody selection systems

Phase	Agent	(Un)Friendly	(Un)Pleasant	(Un)Intelligent	Apathetic/Responsive	(Un)Kind
All	All	3/0.982	4/0.916	4/1.269	4/1.31	4/0.947
All	kotaro	2/1.086	3/1.063	3/1.133	3/1.705	4/1.349
All	screen	3/1.078	3/0.908	3/1.263	3/1.793	4/1.191
All	robot	3/1.335	3/1.063	3/1.12	3/1.771	4/1.124
P_1	All	3/1.022	3/1.22	3/1.188	3/1.662	3/1.174
P_2	All	3/0.739	4/0.491	4/0.485	4/0.585	4/0.625
P_1	kotaro	3/0.906	3/1.194	3/1.262	3/1.335	3/1.226
P_1	screen	3/0.906	3/1.2	2/1.22	3/1.834	4/1.294
P_1	robot	3/1.318	3/1.335	2/1.102	3/1.895	3/1.042
P_2	kotaro	3/0.605	4/0.471	4/0.471	4/0.706	4/0.81
P_2	screen	3/0.769	4/0.385	4/0.533	4/0.584	4/0.456
P_2	robot	4/0.782	4/0.598	4/0.48	4/0.535	4/0.72

perception of intelligence of the agents is rather low during P_1 , except for Kotaro avatar, which has a neutral perception of 3. Such perception, however, increases during P_2 for all agents, which are now considered good. Responsiveness also goes from neutral to good for all agents between the phases. During Phase P_1 , only Plantroid Avatar is considered to have a good kindness score, while others are rather neutral; a perception that also improves to 4 in P_2 .

Thus, once can see that for the proposed adapted Scales, there are many advantages of using semantic speech instead of gibberish speech for ECAs, without any obvious problems. However, although that is not captured by the questionnaires, more than once the GPT-3-based chatbot that was used for powering the conversation of all agents during P_2 and P_3 was very blunt and even rude some times; which never happened for the Gibberish Speech, since it conveys no obvious meaning. Some participants who seemed to be more extroverted had a laugh in such instances, but more serious ones seemed not to enjoy it. This way, for future experiments, in order to better understand what makes some users prefer on communication style to others, it is necessary to obtain a personality profile such as the Big Five personality traits [179].

Such results validate previous understanding that Gibberish speech is less engaging than semantic speech for actual conversations [56] and that humanoid-like ECA tend elicit a higher intelligence expectation in users [180]. Two effects must also be noticed. First, the uncanny valley, that is, that the perception on robots and other smart agents improves as it becomes more human-like when the agents still rank low in anthropomorphism, but, once a certain anthropomorphism is

reached, the perception worsens before it improves again [181]. That happens due to many factors, including the mismatch between the expectations of the capabilities of the agent and the actual performance and that it might look like an ill person, which might instigate the self-preservation instincts of users that will perceive the system as strange or creepy or dangerous and their choices and actions will be judged more harshly [182]. The second effect is that the the context of the interactions also dictate the perceived intelligence scores [183], but since all present agents had the same role of an open-ended conversation agent, the greatest impacting factor on such perception is probably the higher anthropomorphism of Kotaro. The Kotaro agent Avatar had a humanoid-like shape, but did not blink or change its facial expressions, which, in the words of some of the participants, was a little strange; and such factors may have contributed to the higher rating in intelligence, but lower friendliness at the same time.

Regarding the higher friendliness rating of the physical Plantroid robot when compared to both screen-based agents, that was also an expected result, since the agents with higher degrees of embodiment tend to cause higher engagement in users [21, 22, 23, 24, 25] and, since the task at hand was only conversation, it is natural that the robot was considered to be more friendly and open.

7.7 Responses of ranking questionnaires

In order to measure not particular characteristics, but the overall enjoyment of the prosody selection systems and of the agents themselves, the ranking questionnaires were proposed in order to see general preferences. However, as previously mentioned, since some a bug in the program caused some of the volunteers (ID numbers 17, 18, 19, 20, 21 and 27) to not Talk with Kotaro during Phase P_2 , but still ranked it as their favorite agent, calls into question the validity of such a measurement mechanism that depends on the memory of users across 30 exchange with a given agent to rank the prosody selection system and 90 for the agents themselves, might not be a very reliable method. Still, since the data was already obtained, it needs to be analyzed. For the participants that did not talk to Kotaro in P_2 , however, their agent ranking data was disregarded; only the prosody selection system rankings are taken into account for the analysis of the ranking of P_2 . The number of 1st, 2nd and 3rd place rankings of each system, agent and phase combination are shown in Table 7.8; and the ranking of the agents for Phases P_1 and P_2 are shown in Table 7.9.

From such results, it is possible to see that the Constant Prosody System was considered to be the best performing one, except for P_1 when the Plantroid robot agent is considered, where it was considered to have performed slightly worse than the Random prosody selection. Random prosody selection system comes as the second place for P_1 and third place for P_2 . Surprisingly enough, GSIP was considered to yield a better performance for Semantic Speech; a purpose it was not developed for. It did not outperform the random prosody selection system once for gibberish speech, but overtook it in semantic speech. A suspicion is that that is due that GSIP, using the Monte Carlo approach to select good prosody did not guarantee that the best prosody pattern was selected and, at the same time, yielded less variety than the random prosody selection; which might have played a role in it under-performing for gibberish speech, where more variety might have been interesting, since the words have no meaning; but made it perform better than random variations of acoustic prosody for semantic speech, where we expect some prosody patterns to convey certain points, feelings *etc*; but in order to make such claims, more data about the particular opinions of volunteers would have been necessary; and it is a valuable lesson for future experiments.

Regarding the agents themselves, Table 7.9 shows that the robot was the favorite agent of volunteers in Phase P_1 , while the Kotaro avatar was considered to be the better performing agent in Phase P_2 , which is an interesting result, given that it was considered to be the least friendly of the agents. Volunteers were instructed to rank the agents in order of which they enjoyed the most interacting with. and, thus, it was expected that the most friendly one would perform better, but it seems that the higher perceived intelligence or the higher anthropomorphism level played a role in such evaluation. Once again, without knowing the internal decision-making process of volunteers, it is difficult to outline any conclusions.

Once again, it is necessary to remind the reader that, out of all three methods employed for measuring the performance of the systems and agents, the ranking scale is expected to yield the least reliable results, since the ranking only happens some time after the actual interactions have happened, unlike the adapted Godspeed Scale questionnaires that are answered just right after an interaction is finished and the impression estimation from video, which allows estimating the immediate reaction of volunteers to the speeches of the ECAs.

7.8 Impression Estimation from Video

Using the same VGG-16-based and ResNet-18-based neural network architectures used in the Talk to Kotaro experiment, and the considerations exposed in Subsection 7.2.4, the videos of all ECA responses were analyzed and the emotional state e_t for every frame t was estimated, yielding a time series of emotional state of the participants, out of which the average emotion was calculated, yielding the results of Table 7.10. It must be noticed that the first 5 utterances in every conversation of P_1 were fixed utterances; and the first 3 responses of the ECA were fixed for P_2 , the fixed column denotes if only the fixed utterances were used in the comparisons, in order to allow for a more just comparison.

It is possible to see that albeit most of the average emotion is negative, P_1 elicited emotions are more negative than the ones caused during P_2 , since volunteers could understand what the agent is saying and be more engaged in the conversation, instead of trying to figure out what the ECA is trying to convey with its semantic free utterances.

Considering both phases, once again, Constant prosody had the overall best performance, followed by GSIP and then by random prosody selection. Same results extend to Phase P_1 but, surprisingly, GSIP outperforms the Constant prosody and the random prosody methods. That is a very interesting finding, given that the system was not developed to generate adequate prosody for semantic speech but it might be the case that the observation of Section 7.7 might indicate that some variation on prosody is better than none and better than a lot for semantic speech, but more data would be necessary to investigate if that is really the case and, if so, to uncover possible reasons.

Considering only the agents independently from Phase and prosody system, Plantroid robot had the best performance, followed by Kotaro and then by Plantroid Avatar, in line with the results of the Godspeed scale questionnaire for the agents in Phase P_2 , where Plantroid robot outperforms other agents and Kotaro has less variance than Plantroid Avatar in the responses, even if the overall attitude seems to be the same. Same results extend to P_1 and P_2 .

Now, considering each system for every agent in the two phases, we can see that for every interaction, the same patterns mentioned above apply for P_1 , with constant prosody being the best performing and GSIP outperforming random selection for all ECAs. For P_2 , however, for each

agent, Constant prosody seems outperform GSIP by a very small margin, except for Plantroid robot, where GSIP has the best performance.

In order to verify if the same results can be seen on each individual volunteer, the three prosody selection systems are ranked for each participant with usable data according to their performance. Such automatic emotion estimation based ranking is somewhat similar to the ranking questionnaire, but does not suffer from ties and from human memory and *post-hoc* rationalization problems. The results for each participant are shown in table 7.11 and synthesized in Table 7.12.

From the results of the emotion estimation from video ranking analysis, one can see that Constant Prosody has dominated for both Phases, with GSIP coming second, tying with random prosody selection for phase P_1 and outperforming it in P_2 .

Such results are more in line with the analysis performed in on the responses of the previous questionnaires, as calculating the averages of the average emotion of each interaction might not show the actual number of good quality interactions because a very good or bad one can skew the results.

In order to analyze the accuracy of GSIP, it was necessary to compare the impression predicted by the system and calculate Impression caused by all generated gibberish speech. To obtain the caused impression, it is a matter of calculating the difference between the initial displayed emotion and the final emotion and then calculate the norm of the difference between the actual impression vector and the estimated impression. Doing so, the average error was of 0.17; which is way better than the value of 0.27 obtained during validation of the system during its training. What is more impressive is that several nationalities that were not present in the initial Talk to Kotaro experiment were present, but the system did not show an increase on the estimation error.

7.9 Results discussion

From the results obtained by analyzing the response to the questionnaires and the video samples, it is possible to draw some conclusions about the initial research hypotheses.

From the Godspeed Scale Questionnaires alone, there is no support for H_1 – speech generated by the proposed system is perceived as more human-like than speech with constant prosody or with randomly generated prosody. It seems to be on par with prosody generated by other systems

for Gibberish speech; it only seems less responsive, and that perception do not extend to semantic speech. However, regarding how natural the generated utterances are, GSIP performs worse than constant and random prosody selection in P_2 .

Regarding H_2 – speech generated by the proposed system generates more positive impression on volunteers than speech with constant or random prosody patterns; from the results of the video analysis, we can say that for Gibberish speech, the GSIP-based prosody selection system does not generate a better experience than constant prosody, despite generating more positive emotions than Random prosody. However, for semantic speech, GSIP seems to have a better average performance, but, by breaking the performance down to a per-user basis, we get the same results of P_1 .

For Hypothesis H_3 , GSIP successfully generated the desired impression on volunteers, even outperforming the results obtained for training data. Since it is the first system of its kind, it has, by default, the best human impression prediction performance, but it allows for automatically verifying how much an utterance will change the emotional state of humans.

In order for H_4 – test subjects are more lenient with a non-humanoid looking avatar regarding semantic-free speech and eventual bad selection of prosody; to be true, volunteers would have had to be more lenient while evaluating the perceived intelligence, pleasantness and responsiveness of the more animal-like agent and the effect was quite the opposite, volunteers evaluated the pet-like robots worse, showing that the intelligence evaluation is linked to the anthropomorphism degree of the agent.

Since the GSIP-based prosody selection system achieved similar performance for the virtual Plantroid Avatar and for the robot in P_1 and even outperformed the Avatar in P_2 , the data seems to support that the system can be successfully used in Physical robots, there seems to be support for H_5 .

Regarding Hypothesis H_6 , it came as a surprise that, on average, GSIP outperformed constant prosody selection and random in several instances, while it failed to do so for gibberish speech. However, since H_6 presumed that H_n , $n = 1, 2, 3, 4$ would hold true, it is not possible to day that there is support for it; nonetheless, it achieves better results other systems on average, but not on a by-case basis.

However, it is important to notice that such results does not imply that GSIP is useless, but that

using the Monte Carlo prosody generation approach did not yield good results. However, since it has shown to be precise, it can be useful for pre-generating gibberish speech offline and then using it to create a desired impression or, after improving speed, to really test more than only 10 prosody candidates.

Moreover, the system must be retrained with all the new gathered data, which is bound to improve its performance. Such performance, along with the results of the analysis of the Talk to Kotaro experiment data, however, hint towards a personalization route for prosody selection systems, at least for gibberish speech.

7.10 GSIP Experiment Phase 3 - a Study on the Effects of Embodiment Level and Novelty Bias on human Impression of ECAs.

An agent can thus have different levels of physical embodiment [20]; some agents are just text on a display or a voice that speaks to users, while others are robots fully capable of sensing and interacting with the world around them. Since their main function always has a social component, in the sense that they talk with humans, it is important to understand how the level of physical embodiment relates to human perception. If this is properly understood, it is possible to design an agent with sufficient embodiment level, since higher levels of embodiment tend to make an ECA more expensive, but also allow the agent to physically interact with humans and its environment [184].

Studies have already been conducted to investigate the relationship between physical embodiment level and human engagement, human perception of ECA [21, 22, 185, 186], and how well users perform certain tasks when interacting with agents of different embodiment levels [23, 24, 25]. However, there is a possibility that such results are due to novelty preference - the preference for new experiences - which may have played an important role in these results. Previous works, even when acknowledging this possibility, have not investigated the effect of novelty on participants' impressions and preferences.

Experiment phase P_3 was designed to challenge the present understanding that is the higher embodiment level that causes higher engagement and better performance. Thus, it is possible to break H_7 into the following hypotheses.

H_7 : Novelty, instead of embodiment level, is the defining factor in shaping: (a) engagement level, (b) perception and (c) preference elicited by ECAs.

If hypothesis $H_{7,a}$ is true, we expect that the holographic agent will elicit emotions with higher valence in volunteers with experience interacting with robots. If $H_{7,b}$ is true, we expect that there will be a strong negative correlation between participants' level of experience with robots and their impressions of the robot agent; and a positive correlation for the Holographic agent. Finally, if $H_{7,c}$ holds, we expect that for volunteers with higher experience with robots, the holographic agent will be ranked higher than other agents.

7.10.1 Godspeed scale questionnaire responses

In order to understand how physical embodiment and novelty shape human perception of ECAs, and testing hypothesis $H_{7,b}$, we need to compile the responses to the adapted Godspeed Scale questionnaires (results shown in Table 7.13) and calculate the correlation between the responses and the experience with robots across all embodiment levels. We can see that except for the Friendliness of the Holographic agent (3, but only for female participants) and the Responsiveness of the robot agent (5, for male participants), all agents received a good 4 rating in every scale. This is also shown by the lack of significant correlations for the participants as a whole, since there is not much difference in the perception of the characteristics of the agents.

However, if we separate the participants who had more experience with robots, levels 2 and 3, we see such a change in perception. While they maintain a mostly positive perception of the screen agent (median of 4 on each scale), the robot has a lower perception of friendliness (median of 3.5, mode of 3), while the apparent advantageous perception of responsiveness disappears.

To draw further conclusions, it is necessary to calculate Stuart-Kendall's τ_C . Considering all participants as a single group, no relevant correlation could be found between the level of embodiment and all questionnaire responses.

Considering subsets of participants to calculate the correlation between embodiment and volunteer perception, we could find a strong negative correlation between the perception of friendliness for male volunteers who had at least some experience with robots, obtaining a $\tau_C = -0.38$ and a $p - value = 0.046$, that is, the more embodied, the less friendly they considered the agent to be. For male volunteers with little experience with robots (level 1), there was a very strong negative correlation between embodiment level and perceived intelligence, with $\tau_C = -0.88$ and $p - value = 0.053$, indicating they considered the more embodied agents less intelligent.

Regarding the correlation between the experience with robots and the impression of the characteristics of the conversational agents, a strong negative correlation was found between the perceived intelligence of all ECA and the level of experience with robots for male participants, with $\tau_C = -0.38$ and $p - value = 0.002$.

For female participants, a moderate correlation was found between the responsiveness perception of all ECA and experience with robots, with $\tau_C = 0.24$ and $p - value = 0.042$.

When examining the correlation between gender and responses to the questionnaire, no signifi-

cant correlation was found.

7.10.2 Agent ranking responses

In order to objectively know which agent was preferred by research participants, they were asked to rank the three ECA in order of preference. The results of the questionnaire are compiled in Table 7.14, where the results are also broken down by into the different levels of experience with robots. As one can observe, the robot agent received the most first place evaluations and, thus, is considered to be the best performing system considering the responses of all participants. However, if we analyze the relationship with novelty, it seems that it did not perform so well among participants who had previous experience with robots, receiving the same number of first place evaluations as the screen agent for levels 1 and 3 of experience with robot; and actually losing its first place among level 2 participants.

Calculating the Stuart-Kendall's τ_C correlation coefficient between gender, experience level and the classification given to each agent, we have obtained that, for male respondents, there is a strong negative correlation between the experience with robot level and the ranking give to the screen agent, with $\tau_C = -0.47$ and a $p - value = 0.049$. That means that the higher the experience with robots, the higher the likelihood of attributing 1st or 2nd place rankings to the screen agent.

7.10.3 Emotion Analysis from Video

In order to measure the engagement of volunteers while listening to the responses of the ECAs, we estimated their emotional state from their facial expressions, to gauge how much they were enjoying the interaction. Every reaction of the volunteers while listening to the ECAs was recorded using a 4K video camera. Two deep neural networks, VGG-16 and ResNet-18 [163], trained on the AffectNet dataset [164], were used to estimate the emotional state from the participants' facial expressions in Russel's 2D valence-arousal emotion space [68] in each frame of the video recordings. However, such networks only estimate the valence (how positive the displayed emotion is) and arousal (how excited a person is) of the facial expression present in a single frame, resulting in an emotion time series. We have calculated the average emotion for each on the 405 videos

of interactions with the agents and the average valence and arousal values for the interactions for each group of participants are shown in Table 7.15.

From these results it can be seen that the holographic agent had the best performance for all research subjects, since it had the least negative valence average. However, considering only the volunteers who had no previous experience with robots, the robot agent had the best performance; and the performance of the robot is even better in the group that had limited experience with robots. However, for the more experienced participants, it is possible to see a steady decline in the performance of the robot, as well as for the other agents, except for the holographic agent, whose performance improves slightly and then deteriorates again. Finally, if we consider only the subjects who had previous experience with robots, the screen agent had the worst performance, while the holographic agent still had the best.

From the average emotion displayed in the interactions, it is possible to see that for all levels of experience with robots, the holographic agent had a slightly better engagement than the robot agent and performed significantly better than the screen agent. However, for participants with no experience, the robot agent generated a higher engagement and the same tendency can be seen for volunteers with little experience (level 1). However, for participants with level 2 and 3 of experience, we can see that the holographic agent performed significantly better than others.

Since we obtained estimated the average emotion for every interaction, it is possible to calculate the correlation between the level of physical embodiment and the average valence and arousal of elicited by the interaction. No statistically relevant correlation could be found between the experience with robots and average arousal. However, for the average valence, we found a weak negative correlations between valence and the experience level with robots for the screen agent and for the holographic agent ($\tau_c = -0.149$, $p - value = 0.021$ and $\tau_c = -0.112$, $p - value = 0.047$, respectively). Moreover, a moderate correlation was found between valence and the experience with robots for the robot agent ($\tau_c = -0.203$, $p - value = 0.001$).

Now, for the correlation between embodiment level and average valence, we were able to find a weak correlation between embodiment level and valence considering all participants with $\tau_c = 0.142$, $p - value = 0.001$. If we consider only male participants, a weak correlation was found, with $\tau_c = 0.107$, $p - value = 0.051$. For all female participants, we found a moderate correlation with $\tau_c = 0.202$, $p - value = 0.006$.

Considering the correlation between embodiment level and average valence for the participants with a certain experience with robots levels, we found a weak correlation for all participants that had no experience, with a $\tau_c = 0.151$ and a $p - value = 0.005$. However, for participants that had little experience with robots, we were able to find a very strong correlation between the average valence and embodiment level, with a $\tau_c = 0.325$ and $p - value = 0.008$. If we consider volunteers by their gender, we can find significant correlation both for men and women for participants without prior experience with robots ($\tau_c = 0.142$, $p - value = 0.031$ and $\tau_c = 0.192$, $p - value = 0.055$, respectively). For men with other experience levels, no relevant correlation could be found. However, for female volunteers with experience levels of 1, 2 and 3, we found strong correlations, with $\tau_c = 0.345$, $p - value = 0.015$, $\tau_c = -0.627$, $p - value = 0.013$ and $\tau_c = 0.493$, $p - value = 0.051$, respectively.

7.10.4 Discussion

More important than showcasing the obtained data and the calculated Stuart-Kendall's correlation coefficient is interpreting the results of Phase P_3 as to understand their implications regarding hypothesis H_7 and other interesting implications for the design of ECAs, which is done in the present Subsection.

Emotion Analysis and engagement

First of all, regarding $H_{7,a}$ (novelty is the defining factor in shaping engagement level), we can see from the results of the emotional analysis that, while the holographic agent performed better than the robot and screen agents for the all participants and for participants with experience level over 1, which seems to support $H_{7,a}$. It is necessary to notice that, for participants that had no previous experience with robots and for those with limited interactions (levels 0 and 1), the robot agent outperformed the holographic one, suggests that when those agents are somewhat novel, the embodiment level seems to be the most important factor. However, since we calculated the average of the average emotion displayed in each interaction, the results can be skewed by samples with very intense emotional response.

Thus, analyzing the calculated Stuart-Kendall correlation coefficients is a preferable way of

investigating $H_{7,a}$. For the average valence elicited by ECAs in the experiment, we found statistically significant correlations for both the experience level with robots and the embodiment level. However, since the correlations between the experience levels were weak or moderate, showing that while the experience with robot seems to have some effect on the engagement, there is not support for it being the most defining factor in shaping it.

Specially because when we verify the correlation between embodiment level and the average valence, we find weak correlation for all participants, weak for men only and moderate for female volunteers. If we separate the volunteers accordingly to their experience level with robots, we have found a strong correlation for volunteers with limited experience with robots (level 1), but not for higher levels, except for female participants. However it is necessary to notice that levels 2 and 3 of experience with robots had a single female participant in each and, thus, such results shows their personal preferences.

Such results suggest that both the embodiment level and novelty play a role in shaping the valence of the volunteers in their interactions. Curiously enough, no significant correlation was found between arousal and embodiment or experience with robots levels, suggesting that they do not influence how much aroused the participants are from relaxation to excitement in an open conversation setting.

Perception of the characteristics of ECAs

Regarding the Perception of the three ECAs, there was not much difference on the perception between the agents; all agents receive a median rating of 4 for most of their characteristics, which is shown by the lack of significant correlations between embodiment and experience with robot levels and the perceived characteristics of the agents. However, there was a strong negative correlation between the experience with robots of male participants and the perceived intelligence of ECAs, while for female volunteer there was a moderate positive correlation between responsiveness and their level of experience with robots. This way, there is not enough support for hypothesis $H_{7,b}$, since only one characteristic was affected by the experience levels per gender. This is an indication, however, that it does affect the perception, but not as strongly as expected. The negative correlation between intelligence and experience with robots for male volunteers is in line with previous research results, such as [187], which found that research subjects that had previously

interacted with a NAO robot rated its intelligence lower than volunteers that had no previous experience. However, since no participant of our experiment had previously interacted with Plantroid, our results suggest that this effect can carry out across different robots, that is, people that have interacted more with robots are more aware of the limitations of currently available technologies.

Analyzing the most common answer among respondents, we can see that women considered the holographic agent to be of neutral friendliness, but had a positive impression of other agents. Men, on the other hand, had a more polarized opinion, as evidenced by the higher variance among their responses. The robot agent had the most neutral friendliness and kindness responses, but was rated high on the intelligence and responsiveness scale. For all respondents, the robot received the highest score for responsiveness. That is very interesting, since the robot was certainly the least responsive of the agents, since it always had a delay between listening and responding, since it was controlled by the laptop through a socket server and that added a few milliseconds of delay, which should not be noticeable. However, not only was it not noticed, but it caused the opposite impression.

Preference of ECAs

If $H_{7,c}$ was to hold, participants with higher experience with robots were supposed to prefer the holographic agent to the screen and robot agents. However, the fact is that the screen agent was the preferred agent for users with experience levels 1 and 2 and tied with the robot agent for volunteers with the level 3 experience with robots. Considering all participants as a whole, the robot agent was the overall favorite, and the same result extends for the participants that had no previous experience with robots, showing support again to the idea that when both the less and more embodied agents are a novelty, humans will prefer the more embodied one. The only relevant correlation found was between the ranking assigned by male volunteers to the screen agent and previous experience with robots.

Preference for the screen-based agent, is in line with the findings in [186], where the authors found that the robot agent had a harder time expressing emotions than a screen-based one, showing that, for cases where a physical agent is not needed, a screen-based one can be a better choice. Thus, no support was found for $H_{7,c}$.

It is necessary to notice that some volunteers that did not prefer the holographic agent, when

questioned about their choice after the experiment was over, pointed out that they had preferred the robot agent because it was physical, while others said that they had preferred the screen agent because it was bigger than others. Many pointed out that the small size of the looking glass display was a decisive factor for them preferring one of the other agents.

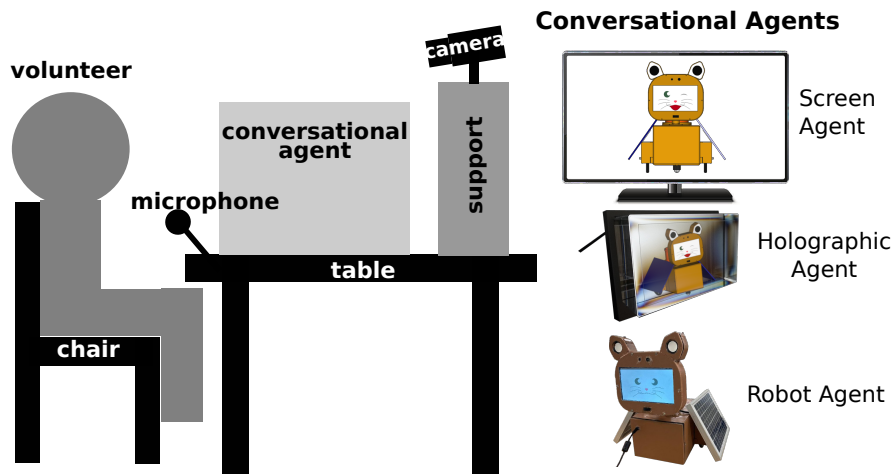


Fig. 7.2: Setup used for the experiment.

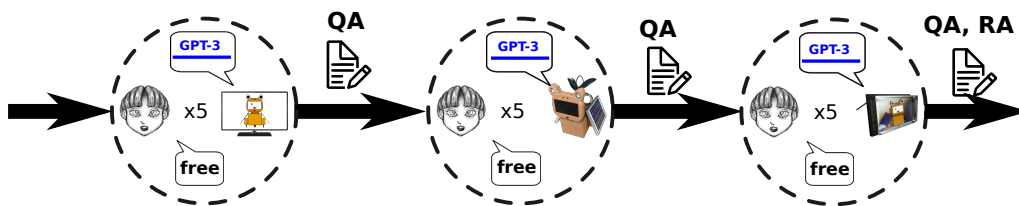


Fig. 7.3: Experiment workflow, where QA is the questionnaire about the last agent and RA stands for Ranking of the agents.

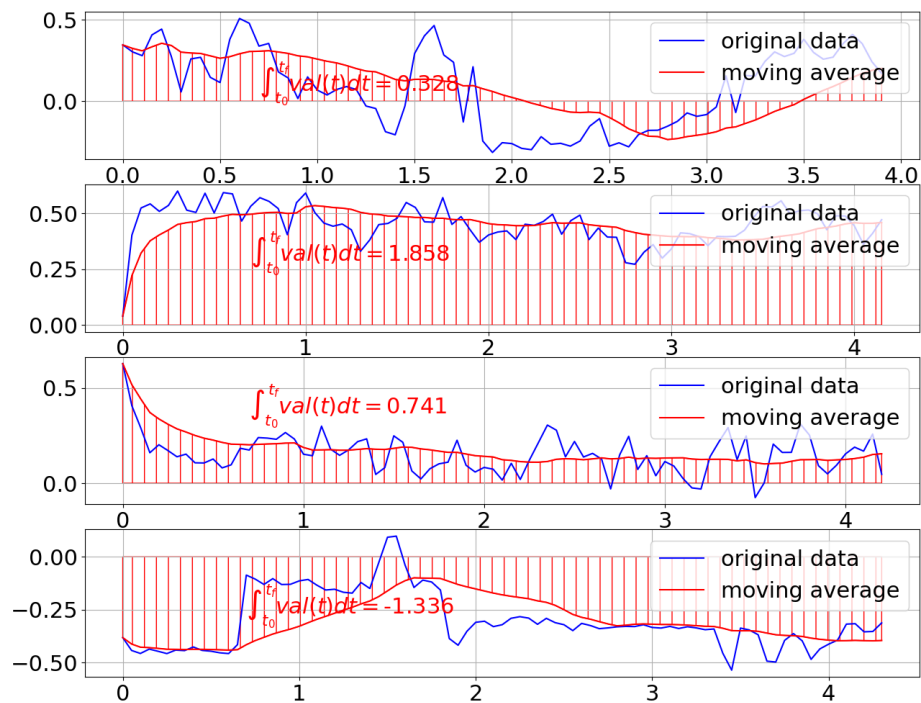


Fig. 7.4: Estimated valence time series (blue), the smoothed data (red line) and the integral of the area under the smoothed emotion curve, calculated to capture the overall emotion of the video.

Table 7.8: Total of first, second and third place classifications each system received considering all possible combinations of Phase and ECAs.

Phase	Agent	System	1 st	2 nd	3 rd
P_1	All	Const	54	20	4
P_1	All	GSIP	28	28	22
P_1	All	Rand	49	25	4
P_2	All	Const	52	13	4
P_2	All	GSIP	21	36	12
P_2	All	Rand	20	33	16
P_1	kotaro	Const	18	7	1
P_1	kotaro	GSIP	8	13	5
P_1	kotaro	Rand	16	8	2
P_1	screen	Const	19	6	1
P_1	screen	GSIP	8	10	8
P_1	screen	Rand	16	9	1
P_1	robot	Const	17	7	2
P_1	robot	GSIP	12	5	9
P_1	robot	Rand	17	8	1
P_2	kotaro	Const	15	3	1
P_2	kotaro	GSIP	6	11	2
P_2	kotaro	Rand	6	9	4
P_2	screen	Const	16	8	1
P_2	screen	GSIP	9	11	5
P_2	screen	Rand	7	12	6
P_2	robot	Const	21	2	2
P_2	robot	GSIP	6	14	5
P_2	robot	Rand	7	12	6

Table 7.9: Total of first, second and third place rankings received by each ECA during Phases $P_{1,2}$.

Phase	Agent	1 st	2 nd	3 rd
P_1	Kotaro	11	8	7
P_1	Screen	13	9	4
P_1	Robot	13	11	2
P_2	Kotaro	11	6	2
P_2	Screen	5	12	2
P_2	Robot	9	4	6

Table 7.10: Average estimated emotion for every system, agent and phase, where fixed denotes that only the fixed Gibberish Speech and questions were analyzed.

Phase	Agent	System	Fixed	Avg (val, aro)	Avg duration
P_1	All	All	No	-0.3171, 0.1385	0.3393
P_2	All	All	No	-0.2897, 0.1245	2.5983
All	All	Const	No	-0.2860, 0.1298	1.2482
All	All	GSIP	No	-0.3028, 0.1319	1.7750
All	All	Random	No	-0.3151, 0.1334	1.5467
All	Robot	All	No	-0.2704, 0.1378	1.5757
All	Screen	All	No	-0.3230, 0.1264	1.6279
All	Kotaro	All	No	-0.3163, 0.1303	1.2371
P_1	All	Const	No	-0.3031, 0.1359	0.2190
P_1	All	GSIP	No	-0.3226, 0.1406	0.4329
P_1	All	Rand	No	-0.3255, 0.1389	0.3654
P_1	Kotaro	All	No	-0.3236, 0.1402	0.3393
P_1	Screen	All	No	-0.3279, 0.1306	0.3374
P_1	Robot	All	No	-0.3171, 0.1385	0.3393
P_1	Kotaro	Const	No	-0.3151, 0.1364	0.2191
P_1	Kotaro	GSIP	No	-0.3172, 0.1342	1.4577
P_1	Kotaro	random	No	-0.3255, 0.1401	0.3606
P_1	Screen	Const	No	-0.3071, 0.1339	0.2185
P_1	Screen	GSIP	No	-0.3383, 0.1323	0.4341
P_1	Screen	random	No	-0.3385, 0.1255	0.3600
P_1	Robot	Const	No	-0.3031, 0.1359	0.2190
P_1	Robot	GSIP	No	-0.3226, 0.1406	0.4329
P_1	Robot	random	No	-0.3255, 0.1389	0.3654
P_1	All	Const	yes	-0.2805, 0.1362	0.1938
P_1	All	GSIP	yes	-0.3051, 0.1423	0.4236
P_1	All	Rand	yes	-0.3202, 0.1413	0.3340
P_1	Kotaro	All	yes	-0.3198, 0.1423	0.2885
P_1	Screen	All	yes	-0.3271, 0.1368	0.2884
P_1	Robot	All	yes	-0.3137, 0.1393	0.2893
P_1	Kotaro	Const	yes	-0.3101, 0.1360	0.1735
P_1	Kotaro	GSIP	yes	-0.3143, 0.1294	1.2989
P_1	Kotaro	random	yes	-0.3137, 0.1475	0.3088
P_1	Screen	Const	yes	-0.3089, 0.1403	0.1720
P_1	Screen	GSIP	yes	-0.3327, 0.1388	0.3909
P_1	Screen	random	yes	-0.3396, 0.1313	0.3030
P_1	Robot	Const	yes	-0.2971, 0.1365	0.1725
P_1	Robot	GSIP	yes	-0.3210, 0.1398	0.3838
P_1	Robot	random	yes	-0.3229, 0.1416	0.3106
P_2	Const	All	no	-0.2833, 0.1231	1.8769
P_2	GSIP	All	no	-0.2828, 0.1231	3.1219
P_2	Rand	All	no	-0.3039, 0.1275	2.8135
P_2	Kotaro	All	no	-0.3071, 0.1179	2.3698
P_2	Screen	All	no	-0.3200, 0.1217	2.7552
P_2	Robot	All	no	-0.2897, 0.1245	2.5983
P_2	Kotaro	Const	no	-0.3009, 0.1186	1.6752
P_2	Kotaro	GSIP	no	-0.3172, 0.1342	1.4577
P_2	Kotaro	random	no	-0.3191, 0.1132	2.6776
P_2	Screen	Const	no	-0.3110, 0.1253	2.0774
P_2	Screen	GSIP	no	-0.3172, 0.1166	3.3045
P_2	Screen	random	no	-0.3325, 0.1228	2.9386
P_2	Robot	Const	no	-0.2833, 0.1231	1.8769
P_2	Robot	GSIP	no	-0.2828, 0.1231	3.1219
P_2	Robot	random	no	-0.3039, 0.1275	2.8135
P_2	All	Const	yes	-0.2698, 0.1264	1.5608
P_2	All	GSIP	yes	-0.2719, 0.1213	2.5655
P_2	All	Rand	yes	-0.2906, 0.1275	2.3870
P_2	Kotaro	All	yes	-0.2783, 0.1130	1.9249
P_2	Screen	All	yes	-0.3121, 0.1308	1.8075
P_2	Robot	All	yes	-0.2676, 0.1247	1.8459
P_2	Kotaro	Const	yes	-0.2890, 0.1151	1.2183
P_2	Kotaro	GSIP	yes	-0.2907, 0.1285	1.2293
P_2	Kotaro	random	yes	-0.2924, 0.1226	2.1419
P_2	PScreen	Const	yes	-0.3318, 0.1419	1.5784
P_2	Screen	GSIP	yes	-0.3139, 0.1215	2.0349
P_2	Screen	random	yes	-0.2864, 0.1259	1.8758
P_2	Robot	Const	yes	-0.2624, 0.1266	1.3679
P_2	Robot	GSIP	yes	-0.2639, 0.1184	2.1884
P_2	Robot	random	yes	-0.2770, 0.1291	2.0217

Table 7.11: Ranking of the performance of each system according to the average emotion estimated from the video camera for all participants who had usable data, where G is GSIP, R is random and C is constant.

ID	Phase	1 st	2 nd	3 rd
5	P_1	R	G	C
5	P_2	C	R	G
6	P_1	C	G	R
6	P_2	G	R	C
7	P_1	R	C	G
7	P_2	G	R	G
8	P_1	C	R	G
8	P_2	G	C	R
9	P_1	C	G	R
9	P_2	C	G	R
10	P_1	G	R	C
10	P_2	C	R	G
11	P_1	C	R	G
11	P_2	G	C	R
12	P_1	R	G	C
12	P_2	R	C	G
13	P_1	G	R	C
13	P_2	R	G	C
14	P_1	G	R	C
14	P_2	C	R	G
15	P_1	C	R	G
15	P_2	C	G	R
16	P_1	R	C	G
16	P_2	G	R	C
17	P_1	R	C	G
17	P_2	G	R	C
18	P_1	C	G	R
18	P_2	G	R	C
19	P_1	C	R	G
19	P_2	C	R	G
20	P_1	R	C	G
20	P_2	C	G	R
21	P_1	C	G	R
21	P_2	R	C	G
22	P_1	G	R	C
22	P_2	G	C	R
23	P_1	C	G	R
23	P_2	C	G	R
24	P_1	G	C	R
24	P_2	R	C	G
25	P_1	C	G	R
25	P_2	C	G	R
26	P_1	R	G	C
26	P_2	C	R	G
27	P_1	G	C	R
27	P_2	C	R	G
28	P_1	G	R	C
28	P_2	C	R	G

Table 7.12: Total of first, second and third place rankings for each user obtained through the video estimation analysis.

P_1	1 st	2 nd	3 rd	P_2	1 st	2 nd	3 rd
GSIP	7	9	8		8	6	11
Random	7	9	8		4	1	8
Constant	10	6	8		12	6	5

Table 7.13: Results of the adapted Godspeed scale questionnaire (median/mode/variance), where A=all, F=female, M=male, R=robot, H=holographic and S=screen.

		(Un)Friendly	(Un)Pleasant	(Un)Intelligent	Apathetic/ Responsive	(Un)Kind
A	R	4/4/0.76	4/4/0.46	4/4/0.94	4/5/0.6	4/4/0.58
	H	4/4/0.66	4/4/0.62	4/4/0.72	4/4/0.67	4/4/0.76
	S	4/4/0.85	4/4/0.42	4/4/0.5	4/4/0.51	4/4/0.52
F	R	4/4/0.68	4/4/0.46	4/4/0.53	4/4/0.53	4/4/0.53
	H	3/3/0.88	4/4/0.59	4/4/0.59	4/4/0.53	4/4/1.08
	S	4/4/0.95	4/4/0.53	4/4/0.38	4/5/0.68	4/4/0.69
M	R	4/3/0.91	4/4/0.5	4/5/1.47	5/5/0.73	4/3/0.64
	H	4/4/0.41	4/4/0.64	4/3/0.9	4/4/0.83	4/4/0.47
	S	4/4/0.81	4/4/0.33	4/4/0.64	4/4/0.36	4/4/0.36

Table 7.14: Responses of the ranking of the agents according to the experience with robots of volunteers.

Everyone	Robot	Holographic	Screen
1st	14	9	12
2nd	9	8	9
3rd	3	9	4
Experienced-3			
1st	3	2	3
2nd	1	1	1
3rd	0	1	0
Intermediate-2			
1st	1	0	2
2nd	1	2	0
3rd	0	0	0
Beginner-1			
1st	2	2	2
2nd	1	2	2
3rd	2	1	1
No experience-0			
1st	8	5	5
2nd	6	3	7
3rd	1	7	3

Table 7.15: Averages of the average emotional response elicited by interacting with each ECA in terms of valence and arousal for each level of experience with robots.

Exp lvl	Screen	agent	Hologram	agent	Robot	agent
	valence	arousal	valence	arousal	valence	arousal
All	-0.31/0.04	0.13/0.01	-0.23/0.05	0.13/0.01	-0.24/0.06	0.13/0.01
0	-0.29/0.04	0.14/0.02	-0.21/0.05	0.12/0.01	-0.2/0.06	0.13/0.01
1	-0.3/0.05	0.15/0.01	-0.33/0.01	0.24/0.01	-0.16/0.06	0.15/0.01
2	-0.39/0.01	0.14/0.0	-0.19/0.06	0.14/0.0	-0.39/0.01	0.14/0.0
3	-0.36/0.03	0.09/0.01	-0.28/0.07	0.09/0.01	-0.4/0.04	0.11/0.01

第8章

Development of the new Social Plantroid

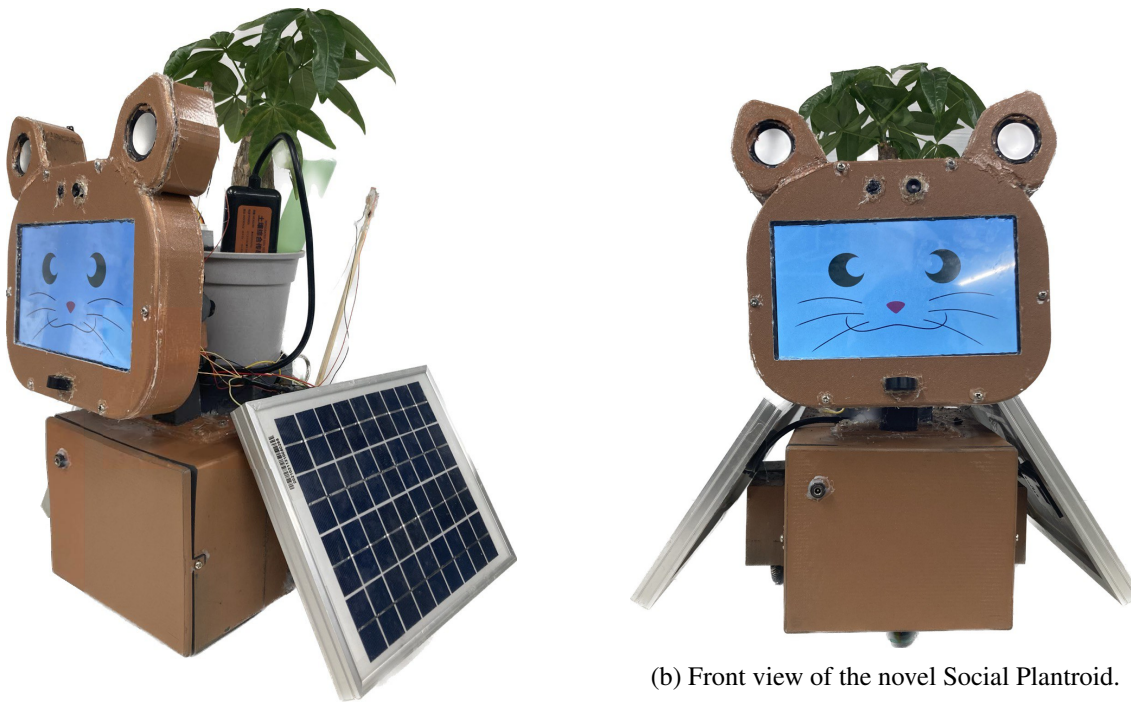
8.1 Introduction

In the Agriculture 4.0 paradigm, the latest technological advances in fields such as internet of things , big data, machine learning, remote sensing and precision farming are put together in order to optimize crop yield and quality, while minimizing the environmental impact, costs and intensiveness of labor ([28]). Moreover, given the current trend of urbanization ([29, 30]) and that many countries have an ageing population ([31]), the reduction of the available workforce and of the average farm was only a natural, albeit perilous, outcome ([28, 29, 32]). In that sense, robotics is expected to play a central role in future of farming due to its resource saving, precision improving and labor saving potential ([28]). Interest on agrobots (agricultural robot) research has, thus, only grown in the last few decades ([33]).

With the reduction of available farmland, greenhouse farming ([29, 33]) and urban agriculture ([34, 32, 35]) appear as very labor intensive solutions, which require precise resource management. Research on IoT, big data machine learning, AI-assisted decision-making systems address the resource management aspect ([36]). The original Plantroid (plant droid) research ([26, 37]), and by extension this present work, are inserted in the corpus of robotics-based labor-saving solutions research. However, whereas previously developed Plantroids , hereby referred to as Plantroid Omni ([26]) (shown in Figure 1.1a) and Plantroid mini ([37]) (shown in Figure 1.1b) only address the labor intensive problem of carrying plants into and out of sunlight in smart-greenhouses and plant factories; the novel Plantroid v.3 (shown in Figure 1.1c) also takes care of monitoring the soil of the plant, information management and communication.

The last function of the novel Plantroid acknowledges the fact that, while Robots and AI might substitute human labor in certain conditions ([38, 39]), it is not expected to happen in the near future and, thus, robots are expected to work side-by-side with human workers ([40]). This way, the robot needs to be able to competently communicate with workers, which requires understanding human verbal and non-verbal communication ([41, 42]). Its pet-like appearance, as seen in Figure 8.1 was chosen to make the robot appealing ([43]) for home-owners who might want a house companion that also helps taking care of potted plants.

Previous Plantroid versions required an external camera for environment navigation and, most importantly, for performing its main task of finding sunlight or shadow, accordingly to the need



(a) 3/4 view of the novel Social Plantroid.

(b) Front view of the novel Social Plantroid.

Fig. 8.1: Novel Social Plantroid.

of the plants they carried. The novel Social Plantroid has two cameras, one gray scale OMRON B5T-007001-010 ([44]), used for human detection, emotion recognition and sunlight detection and an Adafruit MLX90640 IR Thermal Camera ([45]) for sunlight detection.

This way, the novel Social Plantroid was developed to be a Human-Robot Interaction research platform and an agrobot research platform. It addresses the problem that doomed many social robots to fail as products: the lack of perceived utility by customers ([46]). It is, to the knowledge of the authors, the first open-source agricultural robot with a social function. Another novelty presented in this paper is a simple, but effective, sunlight-seeking algorithm which requires no external cameras.

8.2 A Novel Plantroid

The development of Social Plantroid expands the capabilities of the other Plantroid versions in the navigation, vision and social aspects, albeit the swarm capabilities are severely reduced

when compared to Plantroid mini. It was completely built from the ground up without taking any design elements from its predecessors. This way, the developed systems can be split first in hardware (Subsection 8.2.1) and software (Subsection 8.2.2) categories and those categories can be subsequently subdivided into subcategories.

8.2.1 Hardware

The hardware of the Social Plantroid was designed with simplicity and sturdiness in mind, allowing easy reproduction by users and other researchers who possess a 3D-printer, while enabling it to carry heavy loads while being water resistant. It was designed to be water resistant because accidents are expected to happen while the the plant carried by Social Plantroid is being watered. Moreover, if the robot is being used outside, it may also face rain or move over water puddles. The development of Plantroid was done with modularity in mind, so, while the mechanical components (presented in Subsubsection 8.2.1) were developed to accommodate the embedded electronics, the board electronic components (presented in Subsubsection 8.2.1) were chosen to meet the necessities of the mechanical systems.

Mechanical Systems

The Mechanical systems of Social Plantroid consist of the structures of the robot which are not embedded electronics themselves, being responsible for allowing the robot to navigate, to protect the electronics from environmental dangers, to carry plants and to interact socially with humans. Such system, subdivided into social, navigation and support components is show in Figure 8.2.

The Social components are the robot's head (shown in Figure 8.3), which houses its audiovisual components and presents an animal-like appearance, in order to be more appealing to users, and an 1 degree of freedom (DOF) neck (shown in Figure 8.4), which houses a servo, supports the head and allows many wires to pass from the inside of the robot's body to its head.

The head consists of 8 parts: the frontal face plate (i), two frontal ear plates (ii,iii), the main head back (iv), the back part of the ears (v, vi) and the two head supporting structures (vii, viii). The ears and the remainder of the head were designed as separate structures due to the maximum part size that the available 3D printers could print. If a large enough 3D printer is available, they can be printed together.

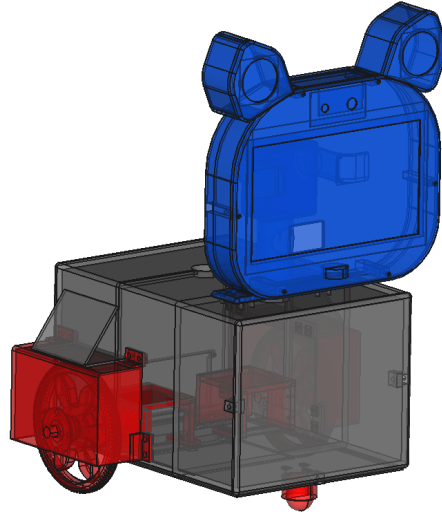


Fig. 8.2: Social Plantroid 's mechanical systems: social components highlighted in blue, navigation components highlighted in red and support components highlighted in grey.

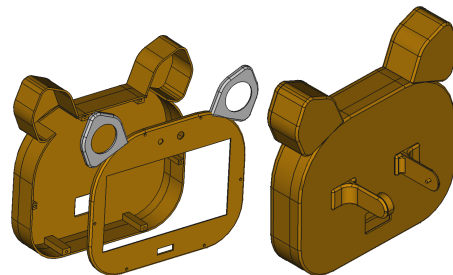


Fig. 8.3: Social Plantroid 's Head frontal and back views.

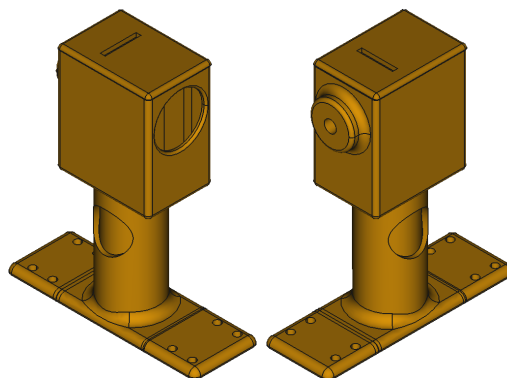


Fig. 8.4: Social Plantroid 's Neck Left and Right views.



Fig. 8.5: Social Plantroid 's Left servomotor holder.

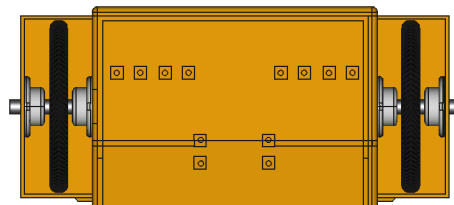


Fig. 8.6: The double bearing design of the Social Plantroid .

The Neck consists of 3 parts: base plate & neck (i), servo case (ii) and servo cover (iii). the servo case was printed separately from the base plate & neck for the sake of saving support material, but otherwise can be printed as a single entity. Moreover, if other researchers or users desire to use a different servo, the system becomes easier to modify.

Regarding the navigation components, this classification is somewhat arbitrary, because without the main body, those parts would be unable to make the robot move. However, since the robot could still exist as an static entity without them, it somewhat be fair to classify them as such. They consist of servomotor supports, bearings, bearing supports, steel axles, wheels, wheel covers and a ball caster.

The servomotor supports (shown in Figure 8.5)are responsible for securely keeping the smart servos who drive the wheels of the Social Plantroid attached to its body, while also serving as support for the lead-acid battery which provides the energy necessary to run the embedded electronics.

The bearings are used in order to reduce friction between the axles and other navigation components, which extends the life of the smart servomotors that drive Social Plantroid . A pair of 8mm ball bearings are used for each wheel. One is held outside the robot's body by a bearing support piece and the other is held to the respective wheel cover by another bearing support. Such double bearing architecture, shown in Figure 8.6 was designed in order to offer greater support to the axles and, consequently, to the wheels.



Fig. 8.7: Social Plantroid 's wheel axle.

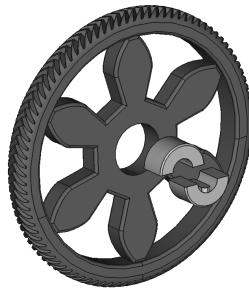


Fig. 8.8: One of the wheels of the Social Plantroid and its components.

The axles themselves (shown in Figure 8.7) consist of a pair of 80mm long stainless steel shafts whose diameter varies along its body. The initial diameter is of 3mm and the, after 20mm, the diameter becomes 10mm. There is also a 10mm long and 2mm deep key-hole in its middle, in order to prevent a wheel with a key from moving from its intended position.

The wheels of the Social Plantroid (shown in Figure 8.8) have a 90mm diameter and are 10mm thick. They were 3D-printed in tough PLA and, thus, works quite well in rough surfaces with small irregularities. However, for very smooth surfaces, it is better to 3D-print the external portion of the wheel in rubber or other material with a higher friction coefficient.

A wheel consists of three parts: the wheel itself, the key axle support and the bottom axle support. After the axle supports are glued to and axle with the key inserted into the key hole of the axle, the supports are glued to the wheel, which cannot slip anymore. If the tolerance between the parts is sufficiently small, no glue is necessary, but printing the parts with higher tolerance allows for easier assembly.

Above and besides the wheels are the wheel covers which are responsible for three functions: supporting the wheels by holding the server, supporting the solar-panel holders and protecting the whole system from any dropped water. Finally, a ball caster is used for allowing the Social Plantroid to navigate with only two servomotors providing differential drive. A commercially available ball caster can be used or 3D printed, accordingly to the resources available.

Joining all the other systems and protecting the embedded electronics, the support system con-

sists of 7 components: three body sections (i, ii, iii); a front cover plate(iv) and a back cover plate (v) and two solar panel holders (vi, vii).

The body of the robot was designed as a box-like structure in order to provide a good and stable surface in which a plant can be safely carried, while providing maximum internal volume for the embedded electronics and other mechanical components. It is divided in 3 sections due to the maximum height that available 3D-printers are able to print. The body of the robot has many holes for bolts, nuts and recessions so the servo holders, front and back covers and wheels covers can be held at their intended positions.

The robot's body features walls with a thickness of 5mm. However, it is possible to reduce this thickness to conserve materials. To enable this, the bottom of the body includes borders (shown in Figure 8.6) that facilitate the installation of a metal, wood, or acrylic plate. This plate serves the purpose of preventing the body from twisting.

As sensors are required to monitor the soil of the plant being carried by Plantroid, they need to transmit information to Plantroid's board computer and, while that could be done by Wi-Fi or Bluetooth, the way which saves most energy is by using a physical connection between the board computer and the plant. For that reason, there is a circular hole with 20mm diameter at the top of Social Plantroid's body. Such hole is covered by the pot, making it water resistant when a silicone adhesive is applied.

The front plate is necessary to completely enclose the internal components inside the body of the robot. It is held in place by two screws and, while it is recommended to apply a silicone adhesive, tests under the rain have shown the robot to be water resistant if the tolerance between the front plate and the body is small enough.

The rear plate is used to give access to the battery, power switch and the rear screws of the smart servomotor holders, since it might be difficult to do so once all parts are assembled in place. The part was designed to be smaller in order to be more water resistant than the front cover, but it can be made larger if easy access is considered a priority over water resistance by researchers and users. It is also held in place by two screws and, once again, it is recommended to be sealed with silicone adhesive to enhance the water resistance, but Plantroid has shown to be resistant against rain without such additional protection.

Finally, the solar panel holders were made out of 1mm thick aluminum plates, which were

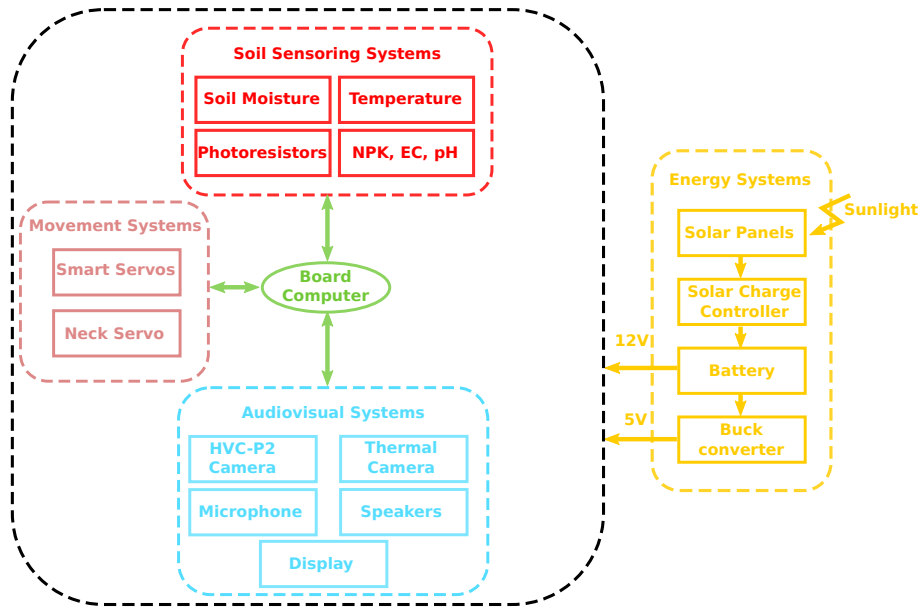


Fig. 8.9: Social Plantroid 's Electronic Systems

bent at a 60° degrees angle. Aluminum was chosen as the material of choice for due to its great corrosion resistance in non-acidic water and to its ductility, since it allows for slight adjustments of the angle of the solar panel being held.

The mechanical design of Social Plantroid still has some room for improvement, but experiments have proven that is a simple, albeit effective design. The many holes allow for easy installation of new accessories for extending Social Plantroid 's capabilities.

Electronic Systems

The embedded electronics of Social Plantroid were selected among off-the-shelf components in order to facilitate reproducibility and increase modularity. The electronic systems are comprised by five subsystems, shown in Figure 8.9: Energy System (i), Board Computer (ii), Audiovisual system (iii), Sensor System (iv) and Movement System (v).

Every electronic system is powered by the energy system, which consists of a 12V 5Ah lead-acid battery, a solar charge controller and a pair of 5W 12v solar panels. The Battery, the solar panels and the step-down buck converter are attached to the solar charge controller (phocos CML 12/24V 20A). The solar charge controller protects the battery from excessive discharge, overheating and

stops charging when the battery is full. However, since the solar panels are low-power, they cannot charge the battery if the robot is in full operation; they only slow down battery depletion rate. The step-down buck-converter provides 5V tension for components that require lower voltage, while components which require 12V, such as the smart servomotors, the integrated soil sensor and the LCD display are directly connected to the input of the buck converter.

Social Plantroid 's Board computer is the brain who coordinates its operation, being responsible for both the agricultural and social sides of the robot. A Raspberry Pi 4 model B with 8GB of RAM is employed for that role. The aforementioned single-board computer was chosen for its compact size, availability of GPIO pins, enough USB2.0 and USB3.0 ports, Bluetooth, Wi-Fi, HDMI port and enough computing power for running its software. Any other single-board computer can be used for its role as long as it provides enough computing power and internet connectivity.

The audiovisual system is responsible for the vision of the robot, capturing speech of users, making sounds and displaying the facial expressions of the robot among other visual information. This system is, thus, the most important one for the social side of the Social Plantroid , besides the Board Computer. A 7-inch hdmi display is used to show a cat-like face and to display other important information, whenever Plantroid is requested to display data about the plant it carries. The loudspeakers are connected to the board computer through USB, as is the microphone. These three components are essential for Social Plantroid 's conversation engine, since they allow it to listen to users, estimate their emotional state from audio, respond and display emotions. However, in order to better estimate the internal state of users, being capable of reading non-verbal cues such as facial expressions is very important. For that, an OMRON B5T-007001-010 OKAO VISION USB camera is connected to the embedded computer. The OKAO vision USB module estimates how positive the emotion being displayed by humans is (valence) and is capable of classifying it into one of five emotions: neutral, happiness, surprise, anger and sadness. Additionally, the grayscale 240×320 image captured by the OKAO vision camera is used for robot navigation, allowing the embedded computer to process it in order to detect sunlit or shadowy spots. Moreover, to help the robot detecting the hottest spots in the floor, a 24×32 Adafruit MLX90640 IRthermal camera was added to the robot; being the only component of the audiovisual system that plays no social role.

By combining the images provided by the audiovisual system cameras, the embedded computer

is capable of determining the location where the robot needs to head to; and the movement system is responsible for doing so. A pair of endless turn Dynamixel AX-12A smart servomotors are used to power the wheels and a 20Kg/cm Lobot servomotor is installed in the neck, allowing the robot to look upwards to hold eye contact with humans and look downwards to detect sunlight and shadow, also providing one additional degree of freedom for bodily language.

Finally, in order to monitor the quality of the soil of the plant currently being carried by the Social Plantroid, many sensors were added to allow the robot to measure some of the most important components that allow predicting the health of the plant. A capacitive soil moisture sensor, a TMP36 temperature sensor, three $10K\Omega$ photoresistors and an integrated smart soil sensor, capable of measuring the concentration of Nitrogen, Phosphorus and Potassium, EC salinity and pH of the soil. Every sensor, except the integrated smart sensor are connected to an Arduino Nano, which connects to the embedded computer through USB. The integrated sensor is connected to the Raspberry Pi through a RS485-USB converter. Since an Arduino Nano cannot measure current, it is necessary to create a voltage divider for the photoresistors, so it can measure the difference of potential between an $1k\Omega$ resistor and the ground. Normally, a $10k\Omega$ would be used, but since the resistance curve of photoresistors are logarithmic, the obtained sensor is not very sensible to more intense sunlight. However, by using a $1k\Omega$ resistor, the sensor becomes more sensible to intense sunlight, in exchange for having low precision in the dark. However, since the Social Plantroid measures light intensity mostly during the day, it does not need high resolution for low-light and, thus, the $1k\Omega$ resistor configuration was chosen.

With the hardware presented in the two previous Subsubsections, a completely assembled Social Plantroid has the performance and characteristics described in Table 8.1

8.2.2 Software

With the hardware developed, it was necessary to make it work as a robot, that is, to be programmable to perform certain tasks. In order to do so, a Raspberry Pi 4B with 8Gb of RAM was chosen to act as the robot's board computer, due to its small size and adequate computational power for the necessary tasks. All code was developed in Python 3 for the ROS2 framework, allowing to use its node structure to obtain greater flexibility, modularity and code integration for

Table 8.1: Performance of the Social Plantroid . All characteristics were measured without any additional.

Metric	Value
Total weight	5.35kg
Maximum linear speed	0.2m/s
Maximum rotational speed	1.9rad/s
Maximum inclination	20°
Maximum movement noise	78dB
Maximum speaking noise	91dB
Battery life (in the dark)	4h
Battery life (strong sunlight)	12h

many sensors.

A ROS2-based framework

ROS 2 was chosen due to its greater speed when compared with previous ROS1 and in order to allow the development of an open-source and modular source code, allowing other researchers to reuse Plantroid’s code as a framework for future social robots since until now there are no other ROS2 social robotics frameworks. Existing frameworks are for ROS1 but since ROS1 will have its development ended, it will not be further update for vulnerabilities, not receive package update for new peripherals, among many other problems.

Since ROS2 maintains a node structure where the different nodes work in parallel and communicate with each other, this allows for greater flexibility and modularity. All of Social Plantroid software was developed with such idea in mind. There is a central main script which manages the routines of Social Plantroid and communicates with every other node whenever necessary, albeit the child nodes also communicate among themselves too. The node structure is as follows:

- MainNode: central node that controls the routines of the robot, taking care of communication, storing and notifying users of problems, taking measurements of the soil parameters from time to time and starting the navigation routines whenever necessary to the plant’s needs.

- **SensorServer**: node responsible for measuring the sunlight levels and the salinity, pH, quantities of Sodium, Potassium and Nitrogen, temperature and soil moisture levels. It can also be expanded to interface with other sensors as required;
- **CameraServer**: node responsible for taking pictures with the thermal and OKAO vision cameras, detecting the position of sunlight through Gabor Kernel Filter, detecting human presence and estimating human emotion;
- **ListenServer**: continuously listen to the environment and performs Voice Activity Detection; sending recorded audio to Google to receive the speech to text, forwarding the text to the Main node which, depending on the contents of the audio, decides to forward it to the chatbot. If a response is warranted, the response will be sent to the GUI node.
- **GUI**: node responsible for controlling Plantroid's facial expressions and of speaking, whenever commanded by the Main node. It implements the dialogue management functions and, thus, it interfaces with the ListenServer to stop Plantroid from listening to its own speech, and interfaces with the GestureServer, which is responsible for controlling Plantroid's bodily language.
- **GestureServer**: node responsible for interacting with the EncoderServer and the NeckServoServer to control the bodily gestures of Plantroid, for example, nodding when saying yes, inclining the head in a bowing manner when meeting someone new, rotating from one side top the othe to express a "no" gestur *etc.*
- **NeckServoServer**: node responsible for controlling the position of the servo in Plantroid's neck through the GPIO pin 11, which is capable of PWM output;
- **EncoderServer**: the encoder server is responsible for far more than just the encoder, it keeps track of the current robot's current pose and issues speed commands to each of the wheels, calculating the speed fro each smart servo from linear and angular speed commands.

There are also helper scripts, such as the `NeuralNavigation.py`, which is started by the `MainNode` whenever it deems necessary for the robot to mve into sunlight and out of it, accordingly to the current time, sunlight and room temperature. Such scripts are as follow:

- `NeuralNavigation.py`: moves Plantroid in a out of sunlight while avoiding obstacles using a VGG-16-based end-to-end visual navigation architecture, which is explained in detail in Subsubsection 8.2.4;
- `ChatBot.py`: implements the NLTK-based chatbot which uses a GPT-J-based Dolly model to give variety to its responses while violating any of the Gricean Maxims, which will be explained in further detail in Subsubsection 8.2.5;
- `ImageProcessing.py`

The `MainNode` implements the routine shown in Figure 8.10

The routine consists of verifying if any humans can be seen or heard. If not, Plantroid will then take care of the plant. It will first check if it is time to check any of the sensors and, if it is, it will check the value and store it. If the value is within the safe range, no notification is stored by `MainNode`. Otherwise, Plantroid will want to tell humans about detected problems, such as dry soil or lack of Potassium in the soil. After that, if the time is between 6h and 18h (or any other time range accordingly to the daylight period at its location), it will check if the plant is receiving enough sunlight. If that is not the case, Plantroid will announce that it is too dark and will use the `NeuralNavigation.py` routine to get into sunlight. Otherwise, it will check the temperature. If it is too hot for the plant species it carries, Plantroid will move into shadow using the `NeuralNavigation.py` routine; and will keep checking if the temperature has reduced. Once it has reduced, it will restart its monitoring cycle.

If Plantroid sensed a human at the beginning, it will try looking at the human to establish eye contact and, if there are any notifications, Plantroid will use its Dialogue Management System to announce it in order to get the human to solve the present problems. If there are no problems, Plantroid will just say a friendly “Hello” to the human, who might or not decide to strike up a conversation. If the human talks first to Plantroid, it will address whatever the human was talking about and, after the conversation topic has finished, Plantroid will announce the current problems that need solving.

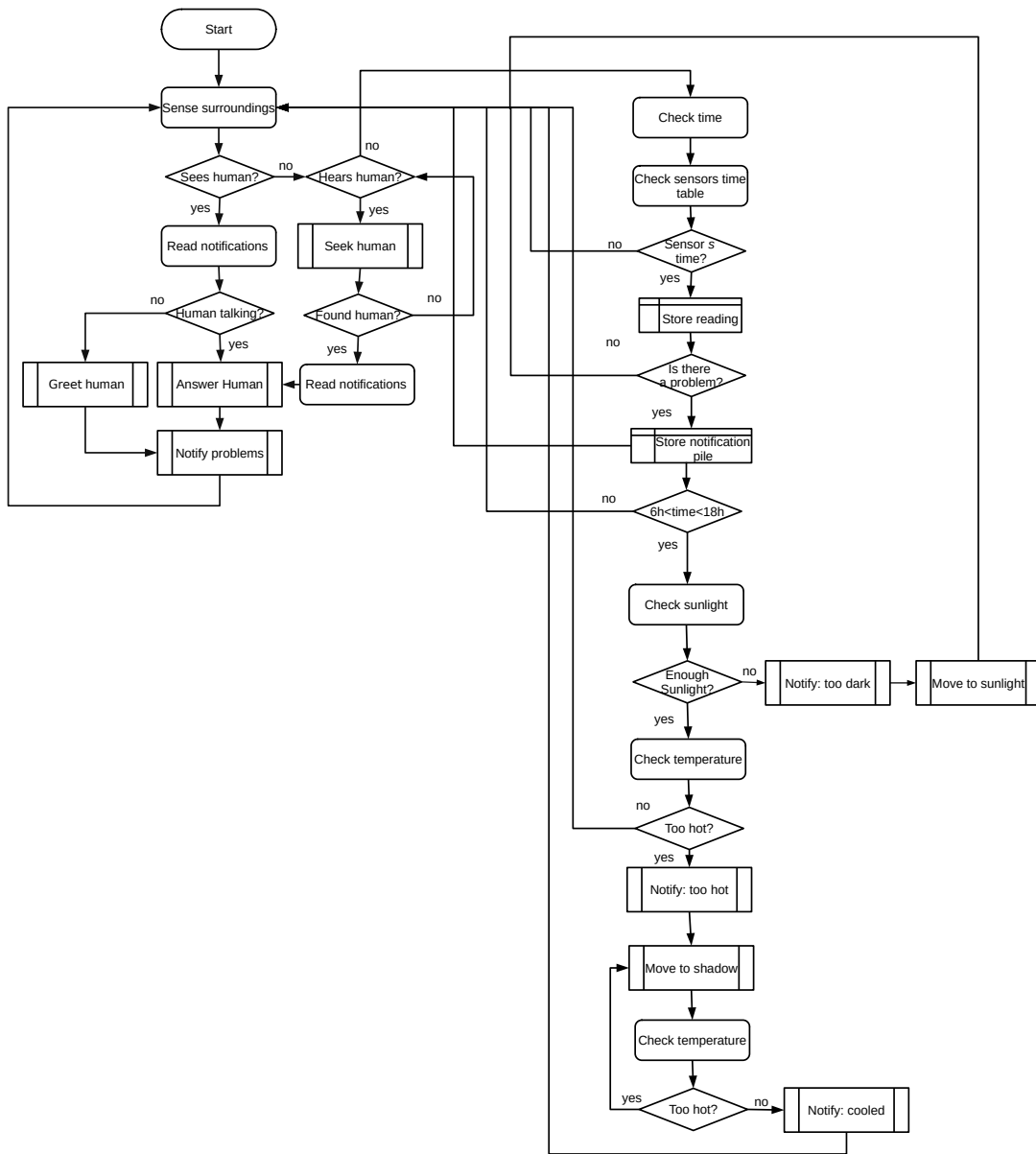


Fig. 8.10: MainNode routine.

8.2.3 Detecting sunlight and shadows

Previous Plantroid robots relied on external cameras strategically placed in the environment for detecting sunlight and deciding where they should go. Trajectory was planned using the APF method, where the intensity of the sunlight was used as the strength of the attractive potential field, while walls and other robots exerted a repellent force.

Such an approach is adequate for smart greenhouses and plant factories, since business owners might be willing to change the environment in order to gain an advantage by automating the caring of plants. However, since the Social Plantroid is a social robot, people would be expected to be less willing to modify their homes and setup a camera to guide the robot. Moreover, even if the novel Plantroid had no agribot functions, it already has a camera for estimating human emotion. Thus, using the same OMRON HVC2-B5T-007001-010 camera for detecting sunlight is a solution that incurs in no extra costs for the robot development allows it to navigate in environments without requiring any adaptations.

Since the OMRON camera outputs $320px \times 240px$ gray scale images, an algorithm for sunlight detection must be able to detect lighter areas over the floor. First, it is necessary to detect the floor itself and, assuming that Plantroid will mostly be deployed in social environments and, thus, is expected to navigate of flat horizontal surfaces most of the time. It is safe, then, to assume that only pixels in the lower half of the image should be considered for the analysis. Moreover, Plantroid has a $24px \times 32px$ thermal camera that can detect warmer spots in the floor, indicating that it is under stronger sunlight. By adjusting the thermal image so it can be overlapped with the gray scale image and converting it into a binary image, we can further reduce the area where it is necessary to detect brighter spots.

Such approach makes it impossible to use the current Social Plantroid near machinery that generate a lot of heat or near fireplaces, since their thermal signatures would convince the robot to go into danger. If that is the case, the warm spot detection part of the proposed sunlight detection approach can be skipped. However, the area over which the sunlit spots must be identified will be larger, increasing computing time.

With the analysis area delimited, the 2D Gabor Kernel filter [188], commonly used for texture segmentation, is used for separating the brightest areas within the same texture. The 2D Gabor Kernel filter parameters that yielded the best sunlight detection results regardless of floor color

are:

- kernel size: 21×21 ;
- orientation of the normal to the parallel stripes σ : $np.pi/6$;
- wavelength λ of the sinusoidal factor: 10;
- spatial aspect ratio γ : 0.5;
- phase offset ψ : 10.

Which delimits brighter areas that have the same texture. The resulting image turned into a binary image using a brightness threshold of 233, that is, any pixel whose value is over 233 is assigned the value 1; 0 otherwise. The resulting binary image is patchy and, thus, it is necessary to perform the dilate operation, where pixels with value 0 neighboring pixels with value 1 become 1. The best performance was obtained by three iterations of the dilation operation with a 7 kernel matrix whose values are all equal to 1.

The resulting image will contain, then, the sunlit areas as 1 and the non-sunlit areas as 0. Since it is a binary image, it is possible to obtain the contour of the white blobs present in it. To find a central point to the obtained contours of the blobs, an algorithm called *polylabel* [189] is employed. *Polylabel* is an algorithm for calculating the pole of inaccessibility [190] of a polygon, that is, the point that is the most distant from its contour. The algorithm can be ran for all blobs but since the plant must be within the sunlit area, *polylabel* is only executed on the blob with the largest area. By selecting the point the furthest away from the contours of the sunlit area, we can ensure that *Plantroid* does not need to move for the longest time possible, since it will take longer for the sunlight to move away from the robot as the sun sets down. Since the field of view of the camera is known (50°) and it is possible to experimentally obtain its focal distance, since the datasheet does not disclose that information, we are able to estimate the distance between the pole of inaccessibility of the largest sunlit area and the robot. For this purpose, we assumed that the floor is horizontal and that there is no lens distortion. which are quite strong assumptions. However, once at least one of the photoresistors of *Plantroid* detects enough sunlight, the navigation is switched from visual navigation to a photoresistor based guidance, where the robot moves trying to

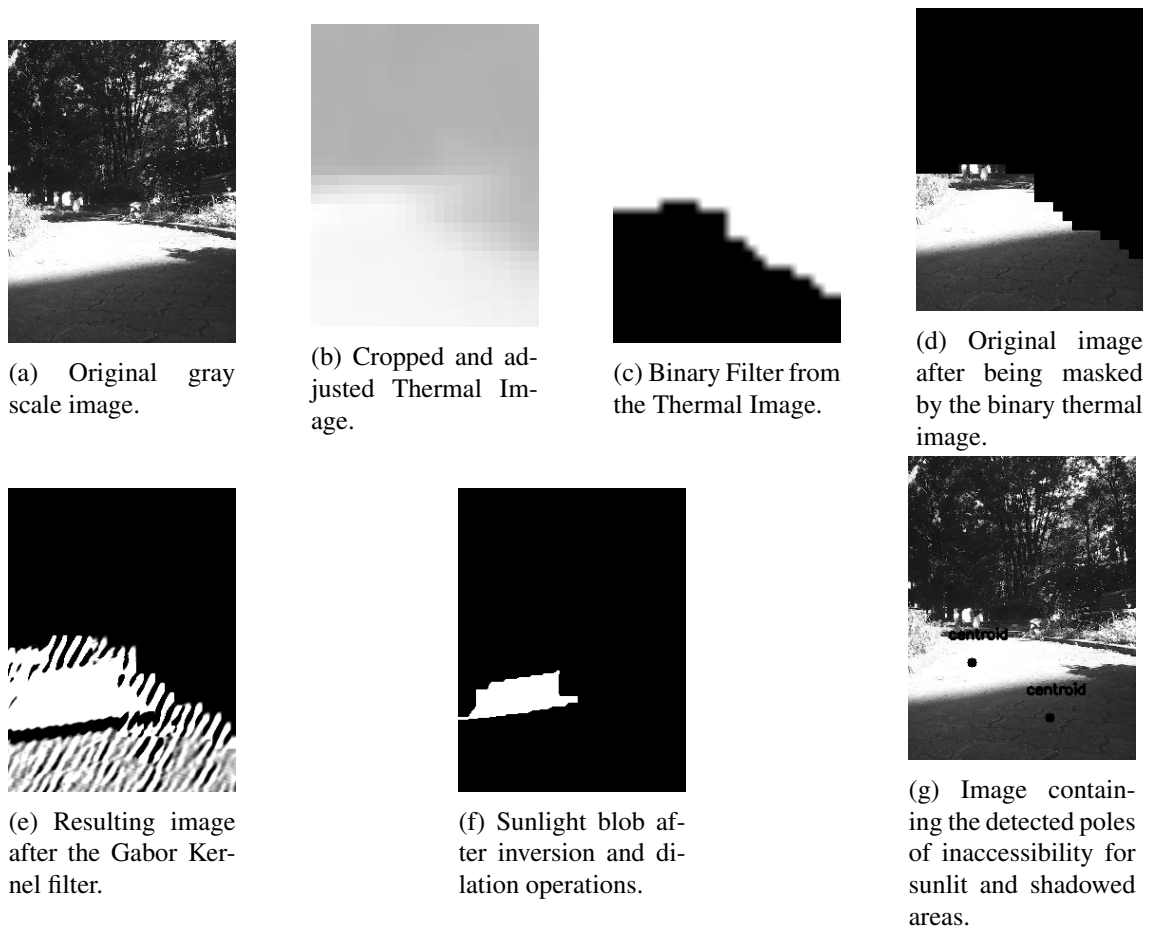


Fig. 8.11: Novel Gabor kernel filter-based algorithm for sunlight detection-step by step.

get all 3 photoresistors under enough sunlight. If the photoresistor that detects the strongest sunlit is the one in the left ear, for example, the robot needs to keep moving forward and turning left, the linear speed determined by the difference between the desired sunlight level and the detected level on the right ear sensor. When both have enough sunlight, the robot moves forward until the tail photoresistor receives enough sunlight. If the first photoresistor to detect enough sunlight is the tail sensor, Plantroid will rotate until one of the ears get enough sunlight and, thus, the behavior switches to the behavior initially described. This way, the estimate of the location of the pole of inaccessibility does not need to be very precise.

A step by step result of the image processing operations described in the Subsection can be seen in Figure 8.11.

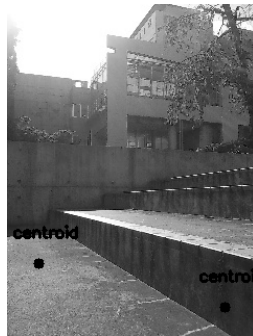


Fig. 8.12: Failure of the shadow detection algorithm.

The same proposed algorithm can be used for detecting shadowed areas, it is only necessary to invert the result of the dilation operation after the Gabor Kernel filter and the results can be seen on the final Subfigure 8.11g.

However, such algorithms might fail in the presence of obstacles, as shown in Figure 8.12, where the detected shadowed area is the side of the step of a ladder. Thus, navigation and obstacle avoidance techniques are necessary, since the proposed sunlight detection algorithm only detects the robot's final destinations. An end-to-end neural network based approach is proposed to solve such problems in Subsection 8.2.4.

8.2.4 Vision-based navigation system

Vision is a powerful sense for navigation – it can be used to achieve the 4 principal tasks of robot navigation: localization, mapping, path planning, and locomotion [47]. Moreover, since most humans rely on vision to navigate in their daily lives, interest in vision-based robot navigation research is natural. This work is inserted in that context, focusing on end-to-end locomotion and obstacle avoidance using monocular gray scale images to estimate the heading direction of a differential drive Plantroid robot.

The main task of a Plantroid robot (Plant + droid) [27] is to seek sunlight or shadow according to the needs of the plant it carries, but the novel model incorporates social aspects into its goals, in order to turn plants into pets. Thus, the environment where the robot navigates changes from smart greenhouses and plant factories to the same places where people live, such as houses, restaurants, and workplaces. To fulfil such requirements, the robot is equipped with a gray scale OMRON B5T

HVC-P2 camera. Even though the proposed VGG-16-based architecture is deployed in Plantroid, it can be extended for other ground robots.

Initially, most approaches for visual navigation were based on Image Processing [47] techniques, but as computing power increased and machine learning matured, machine-learning-based solutions became more present in the field. However, one gap in end-to-end neural-network vision-based navigation research is that no solution was developed to directly learn the behavior derived from using the Artificial Potential Field (APF) [48] method for path planning, the same method used in the previous Plantroid models for seeking sunlight, while avoiding walls and other robots [27].

Vision-based end-to-end methods have the advantage of eliminating the need for robot localization and mapping the environment, generating locomotion decisions by directly sensing the environment, a behavior known as reflex approach [49]. That allows to reduce necessary computation power and reduce the number of necessary sensors, making robots cheaper, lighter, smaller and more energy efficient. Thus, in this work, localization and mapping problems are not addressed; and it assumes that for the task of seeking sunlit areas, Plantroid encoders are precise enough, since the objective destination is an area far larger than the robot itself. Knowing the map is not essential, since the architecture successfully learns how to avoid walls, static and mobile obstacles.

Works [50, 51, 52] have used monocular images to estimate the distance of obstacles from the robot and then used variations of the APF method for trajectory planning. The proposed VGG-16[53] based architecture yields the robot heading directly from images, eliminating the need of running the APF method while achieving comparable performance. APF was chosen as the planning method for training data generation for the proposed architecture because Plantroid's main navigation goal is to move into and out of sunlight while avoiding obstacles. The intensity of the sunlight also translates well into the attractive potential of the robot's goal, as it was done for the previous model, albeit from an external camera. Moreover, its implementation is simple and has many variants. Trajectories planned through APF method are followed through the virtual robot approach, which is also easily implemented.

Using the APF method and the virtual robot approach, over 30h of simulations were run in 3 distinct environments: a house, a cafe, and a meeting room, where the robot navigates from

an initial position $p_i = (x_i, y_i)$ to a final position $p_f = (x_f, y_f)$ while avoiding mobile and static obstacles. Every second, an image is saved, together with the current robot pose (x_r, y_r and θ_r), current destination, and the future heading of the robot obtained through the aforementioned techniques. Generating data, training, and evaluating the navigation architecture in a simulated environment allows for cheaper and faster development since it does not wear out real robots, does not require modifications to the environment and does not need to run in real-time.

Problem definition

This work solves the problem of safely navigating a Plantroid robot r , shown in Figure 8.13 in a social environment Σ_i that has a set of obstacles $O_{\Sigma_i} = O_{s,\Sigma_i} \cup O_{m,\Sigma_i}$, where O_{s,Σ_i} is the subset of static obstacles $o_{s,\Sigma_i,j}$ and O_{m,Σ_i} is the subset of mobile obstacles $o_{m,\Sigma_i,j}$.

The robot state is defined as: $s_r(x_r, y_r, \theta_r, v_{l_r}, v_{\theta_r})$, where x_r, y_r, θ_r are the location and heading of the robot and v_{l_r}, v_{θ_r} are the linear and rotational speeds. Thus, $\dot{x}_r = v_{l_r} \cos(\theta_r)$, $\dot{y}_r = v_{l_r} \sin(\theta_r)$, $\dot{\theta}_r = v_{\theta_r}$.

For an obstacle o_n , its state is defined as $s_{o_n}(x_{o_n}, y_{o_n}, \theta_{o_n})$, where $x_{o_n}, y_{o_n}, \theta_{o_n}$ are position and heading of the obstacle; and $v_{l_{o_n}}, v_{\theta_{o_n}}$ are its linear and rotational speeds. Thus, $\dot{x}_{o_n} = v_{l_{o_n}} \cos(\theta_{o_n})$, $\dot{y}_{o_n} = v_{l_{o_n}} \sin(\theta_{o_n})$, $\dot{\theta}_{o_n} = v_{\theta_{o_n}}$.

The robot navigation problem consists, then, of reaching a final state $s_{r,f}: (x_{r,f}, y_{r,f}, -)$ from an arbitrary initial state $s_{r,i}: (x_{r,i}, y_{r,i}, \theta_{r,i})$ without colliding with obstacles or walls. Currently, $\theta_{r,f}$ is not considered for the final desired state – the robot only needs to reach coordinates $(x_{r,f}, y_{r,f})$. The robot is considered to have reached its final destination if it is within a distance D of the final point, which, for this work is considered to be of $0.1m$.

Plantroid only uses its encoders to obtain its present pose in relation to its origin and the 240×320 px gray scale images from its camera to avoid collisions. In order to accomplish that task, the robot uses the current robot orientation θ_r , the angle θ_l the robot should be heading if it were to go in a straight line to its goal (x_f, y_f) , the distance $d_f = \sqrt{(x_f - x_r)^2 + (y_f - y_r)^2}$ and the camera image I to estimate the next heading angle $\theta_{r,n}$ for a time horizon of T seconds. For this paper, experiments were run considering a time horizon $T = 1s$, but a higher or lower update rate can be used accordingly to the available computing power of the robot.



Fig. 8.13: The New Plantroid Robot.

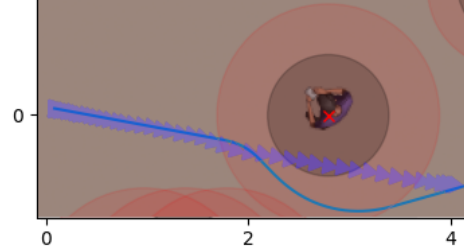


Fig. 8.14: Trajectories of the robot obtained through the artificial potential field method (teal line), the proposed navigation policy P (blue triangle line).

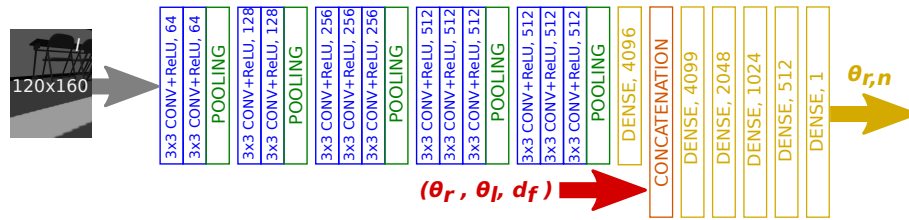


Fig. 8.15: Proposed heading direction estimation architecture.

Proposed architecture

To solve the problem described in Section 8.2.4, an DCNN architecture capable of using the current robot orientation θ_r , angle θ_l , distance d_f and the camera image I (resized to $120 \times 160 \times 1$ due to RAM memory constraints) to estimate the next heading angle $\theta_{r,n}$ for the robot is used. Since two different types of data are being used, two neural networks are used, a VGG-16 DCNN, which receives I as input and a regression neural network, which receives both the output of the VGG-16 network and vector $[\theta_r, \theta_l, d_f]$ as inputs. The regression neural network, then, estimates $\theta_{r,n}$. The layers of the proposed architecture are shown in Figure 8.15.

Data Generation and Neural Network Training

In order to train the proposed architecture, a data set consisting of navigation images and associated data is necessary; and simulation is a great way of generating one, since it does not wear out the robot, does not require modifications to the environment; allows for easy changes on the environment and might even save time since simulations do not need to run in real-time.

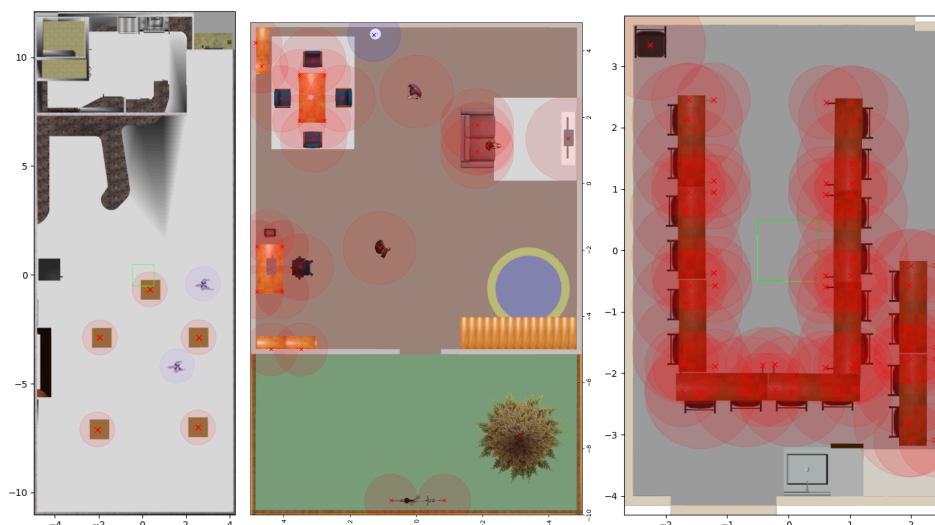


Fig. 8.16: From left to right, simulated environments Σ_1 , Σ_2 and Σ_3 . Static obstacles are highlighted in red and mobile obstacles are highlighted in blue.

Thus, three social environments Σ_1 , Σ_2 and Σ_3 , shown in Figure 8.16, were created in Gazebo Simulator, representing, respectively, a cafe, a house and a meeting room (a virtual version of a real meeting room in Tokyo University of Agriculture and Technology). Those environments have static and mobile obstacles (highlighted in red and blue, respectively in Figure 8.16), with the exception of Σ_3 , which only has static obstacles.

The navigation task is the same as defined in Section 8.2.4, but the navigation goal $(x_f, y_f, -)$ is randomly chosen, while the initial robot pose is either $(0, 0, 0)$ or the final goal position of the previous navigation task. In order to verify if the chosen goal point is reachable, the Artificial Potential Field Method is used and, if after an arbitrary time horizon of 2000 iterations the robot cannot reach the goal point, a new goal is chosen until a reachable destiny is picked. An example of such check can be seen in the teal-colored line trajectory shown in Figure 8.14.

After a destination is selected, the robot navigates to it using the following navigation policy P :

1. If an obstacle or wall is at a distance $d \leq 1.5m$ and within the field of view of the camera of the robot, use the Artificial potential field method to obtain a $\theta_{r,n}$ which avoids obstacles and walls;
2. Otherwise, move in a straight line towards the goal position, that is $\theta_{r,n} = \tan^{-1}((y_f -$

$$y_r)/(x_f - x_r));$$

3. Every second save current camera image I , navigation information θ_r, d_f, θ_l and next robot heading $\theta_{r,n}$.

The navigation policy P considers that the robot cannot dodge obstacles it cannot see unless other sensors are used. To verify which obstacles were inside the robot's field of view and the arbitrary distance d , the virtual Plantroid was equipped with a LiDAR which has the same field of view of the camera, even if the real robot has not such sensor. This was done because it allows for detecting close obstacles in a computationally inexpensive way. An example of a trajectory navigated using this policy can be seen in the blue trajectory present in Figure 8.14.

Data generation

Simulations were ran in all three environments Σ_1, Σ_2 and Σ_3 using navigation policy P during two stages S_1 and S_2 . During S_1 , Plantroid performed navigation tasks in a given environment Σ_i until more than 33,333 pictures were taken for each environment, totaling over 100,000 images. Since less than 25% of the generated data had $|\theta_l - \theta_{r,n}| > 0.01rad$, a second stage S_2 where the images and navigation information are saved only when the robot is close to an obstacle was necessary. Plantroid navigated in all 3 environments once again until no less than 8,900 images per environment were saved, bringing the total saved image number to almost 130,000.

During both stages, the lighting conditions, the position of obstacles and, for Σ_1 , the texture of the ground were periodically changed to enrich the data set in an attempt to give the proposed architecture generalization capabilities. With the combined data set, it became possible to train the proposed architecture.

Model Training

The proposed DCNN-based architecture was trained first using the generated data set it learned how to navigate from the starting position of the task to the goal point in a straight line and could successfully learn obstacle-avoiding behavior which is similar to the results presented by the navigation policy P .

In order to train the model, Adam [191], with a Learning rate of 10^{-3} and a Learning rate decay of 0 was used. The chosen batch size was of 32 images and the training was performed in 20 Epochs.

8.2.5 Experiments and Results

To validate the proposed architecture, two series of experiments were ran. First, in Subsection 8.2.5, several experiments were run in the virtual environments Σ_1, Σ_2 and Σ_3 , out of which 6 will be showcased. They consisted of simulating pairs of navigation tasks in each one of the simulated environments; the first task is executed using navigation policy P and the second is performed using the proposed architecture. For every figure in this Subsection 8.2.5, trajectories obtained through P are shown in blue, while the trajectories obtained by the proposed architecture are shown in yellow.

In each experiment, resulting trajectories are compared using the discrete Fréchet distance metric [192]. The aforementioned metric represents the maximum deviation distance between a trajectory and a given reference curve. It is obtained by calculating the minimum distance between every point of the trajectory and the reference curve, and then selecting the maximum distance among these.

Moreover, to validate the effectiveness of the proposed architecture in real-world settings, experiments were conducted in the real version of environment Σ_3 . These experiments demonstrated the architecture's capability to transfer simulated data-learned behaviors to real-world scenarios in a sim-to-real manner.

Simulated experiments

For the environment Σ_1 , two navigation tasks were performed, one in which the robot needed to avoid two tables in their original location (which have one leg and the robot can go under it) and a second task where the robot must avoid a sequence of 4 tables, none in their original location. For the first task, shown in the left side of Figure 8.17, both trajectories are quite close, with a maximum deviation of $0.233m$ and an average deviation of $0.072m$. In the second task, shown in the right side of Figure 8.17, the robot presented a greater reaction to the presence of the table for

the navigation Policy P , causing a larger Fréchet distance of $0.574m$, with an average deviation of $0.142m$. However, it is possible to see that the trajectory obtained with the proposed architecture was very close to the one obtained by P from the third table onward.

Experiment 3, whose trajectories can be seen in the left side of Figure 8.18, was executed in Σ_2 . The robot moved from its starting position to an arbitrary point $(-4, 2)$, where it should avoid a woman standing in its way. The maximum deviation presented was of $0.338m$, with an average deviation of $0.102m$.

In experiment 4, also executed in Σ_2 , the robot had to dodge a tree, moving from $(-9.5, -3)$ to $(-6, -3)$. The interesting result is that despite repeating the experiment multiple times, the proposed architecture and navigation policy P avoid the tree by going in different directions, yielding a Fréchet distance of $0.621m$ with an average distance of $0.2m$, the largest in the experiments.

Environment Σ_3 has proven to be very hard for navigation, given the large quantity of chairs and tables, whose thin legs easily escape the field of view of the camera once Plantroid moves. Experiment 5 (whose results are show in the left side of Figure 8.19) is a simple task of avoiding the legs of the last table from the bottom row of the meeting room and both P and the proposed architecture achieve the goal, with a Fréchet distance of $0.245m$ with an average distance of $0.101m$. Policy P once more is more reactive to the presence of obstacles than the proposed architecture, but both trajectories are not so distant from each other.

Finally, in experiment 6, two of the chairs were moved to the middle of the room, making it harder for the robot to move around. Policy P fails the navigation task, as one can see by the short blue trajectory in the right side of Figure 8.19). The robot becomes trapped in a local minimum and cannot reach its destination, while the proposed architecture successfully avoids both chairs and reaches its destination, showing that the proposed architecture is more robust than the traditional APF method.

The simulation results were quite satisfactory; the Social Plantroid robot has a width of $0.4m$ and, thus, its body would be over the trajectory obtained by using P most of the time. Moreover, since Plantroid's navigation task does not require extreme precision, it is an acceptable result, specially because the proposed architecture is more robust than the APF method, since it does not get trapped in local minima. All experiments were executed in distinct lightning conditions and the results were virtually identical, showing that the proposed architecture is robust to changes

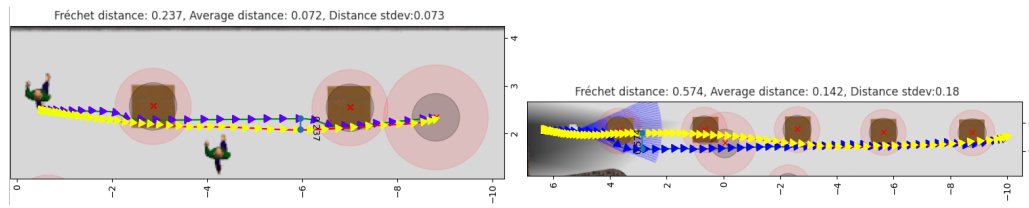


Fig. 8.17: Trajectories of the robot obtained through the policy P proposed architecture in experiments 1(left) and 2(right).

in illumination. Additionally, experiments were ran with obstacles in distinct positions from the original setup and the architecture managed to avoid them.

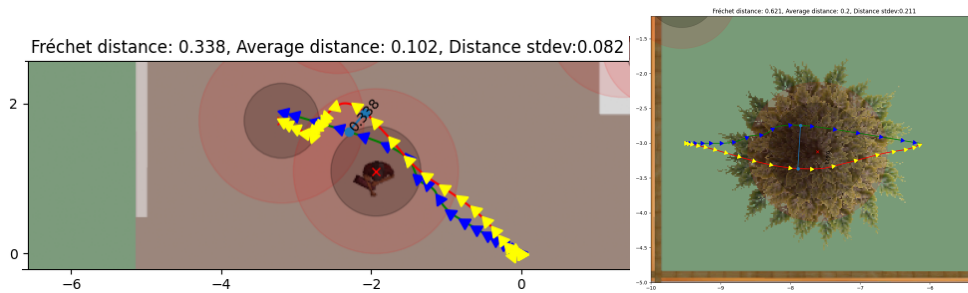


Fig. 8.18: Trajectories of the robot obtained through the proposed navigation policy P proposed architecture in experiments 3(left) and 4(right).

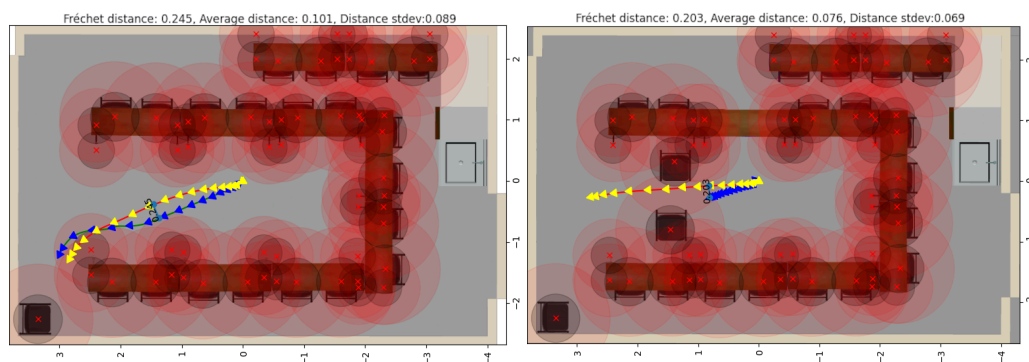


Fig. 8.19: Trajectories of the robot obtained through the proposed navigation policy P proposed architecture in experiments 5(left) and 6(right).

Real scenario experiments

In order to further test the performance of the proposed architecture and the viability of using data obtained through simulation in rather simplistic Gazebo environments, we have deployed the trained neural network in the physical Social Plantroid robot. Since environment Σ_3 is based in a meeting room in Tokyo University of Agriculture and Technology, the experiments were ran in the aforementioned room.

In the real experiments, the navigation task was to move the robot into an area with stronger light, one Plantroid's actual tasks. Figure 8.20 showcase successful experiments were Plantroid achieved the goal in all tries without problems. Figure 8.21 showcase problematic experiments where the limitations of vision-based end-to-end navigation were evident, since the robot tried to avoid the obstacles, but ultimately collided against them once they were out of its field of view.

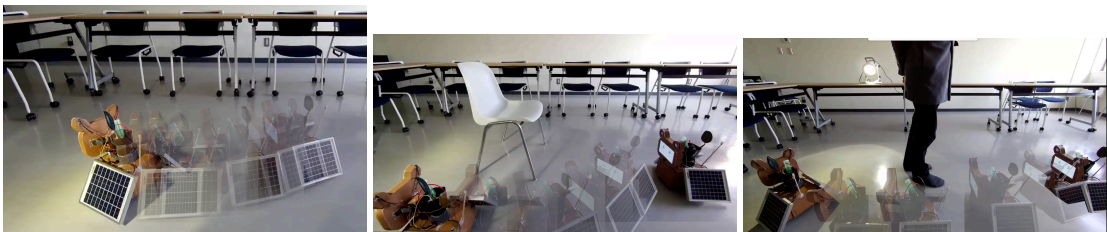


Fig. 8.20: Trajectories of the real Plantroid without obstacles (left), avoiding a chair(center) and dodging a person (right).



Fig. 8.21: Problematic results, many tries were necessary until Plantroid could avoid the legs of the table (left), could not dodge the chairs after losing sight of its legs (center and right).

The experiments in the real scenario can be seen in this video., but the results show that the architecture can actually be deployed in a real case scenario and shows that the behavior learned

form simulation is transferred for the real robot.

Sensoring system

By storing the sensor data in an SQLite database and leveraging the plant growth models described in Section 2.3, Plantroid is able to learn many unknown parameter of the carried plant and, with enough data, is able to predict when the plant will next need water and fertilization. It is also capable of predicting the long term impact on the health of the plant if any deficiency is kept for a longer period of time.

Dialogue Management system

With the advent of large language models (LLM) [193], particularly Generative Pre-Trained (GPT) [194] models, the conversational capabilities of artificial intelligence has tremendously increased and, thus, dialogue management architectures must take advantage of the novel expanded capabilities, but also take care with new problems. Even though there was no unified framework or clear paradigms for the development of such systems [195], how researcher and developers create dialogue management systems is bound to change a lot. For example, OpenAI's GPT-3 davinci model was able to hold open-ended conversations with research volunteers in the GSIP experiment. The problem of using such stock chatbot models is that they are user driven, that is, the human user needs to take initiative. Thus, even with the expanded capabilities of the more recent models, Dialogue managers are still necessary.

Since Plantroid is not expected to engage in extremely complex conversations, as its social presence should be more akin to a very friendly pet who also helps you take care of plants, a smaller, yet powerful GPT-J-based Dolly LLM [196, 197] is used in a similar fashion to how the GPT-2 model is used in [195], where it is responsible for translating precise information into natural language. The problem of current GPT models is that the Transformers they use as their building blocks only learn probability distribution association between words and, thus, cannot guarantee that they will answer the truth, specially when it is able to output associations with lower probabilities. [198, 199]. While this allows the models to sound more creative and interesting to interact with, it also makes them more dangerous, as they will confidently tell non-truths and make

up information on the flight, instead of admitting not knowing something. The models are not lying because there is no malice or deceptive intent behind, thus, calling it hallucination is adequate. In [195], the authors found a way of leveraging the creativity of the models and giving accurate information, by creating an intermediate language that the manager uses, which is translated into natural language by the GPT-2 model; which allows the developed system to be used in many languages, although it seems to have better performance in some of them.

In order to allow Plantroid to work in several languages and to respect the Maxim of Quality (do not say what you believe to be false; and do not say that for which you lack adequate evidence.) and Quantity (Contributions to the conversation are as informative as required; do not make your contributions more informative than is required.), at which GPT models tend to not to be good at respecting, Plantroid's Dialogue Manager uses a similar approach, where there are two chatbots. First chatbot is a simple Python NLTK rule-based chatbot which was handcrafted to give very exact and precise information. The NLTK chatbot answers with requests to the MainNode, such as to measure the salinity of the soil, produce a complete checkup of the plant, move the robot, alter its facial expression or precisely repeat what a user has commanded it to repeat. It is also capable of requesting that information is fetched from trusted source on the internet, when users make questions like "What is X?" or "What is the definition of Y". Currently a dictionary is used for the definition of words and Wikipedia is used in the "what is X?" questions, even though the open encyclopedia may contain falsehoods. However, since this is done to demonstrate the capacities of the system, it is deemed as appropriate. Once the NLTK chatbots requests have been fulfilled by the MainNode, if necessary, the information obtained by the managing script will be sent to the GPT-J Dolly model, which will convert the information into natural language while giving it variety and keeping the conversation interesting. If the NLTK chatbot has no answer for what a user has said or asked, the resulting text of the speech to text operation will be forwarded to the GPT-J Dolly model and its output will be synthesized by espeak.

I'm sorry, I do not know. Plantroid, currently, is capable of admitting that it does not know an answer for a question. It is important to do so, as the robot might be consulted for information that it should not be consulted, such as medical, financial or political advice. If the robot tries to fetch some information and is unable to find it from a reputable source, instead of forwarding the

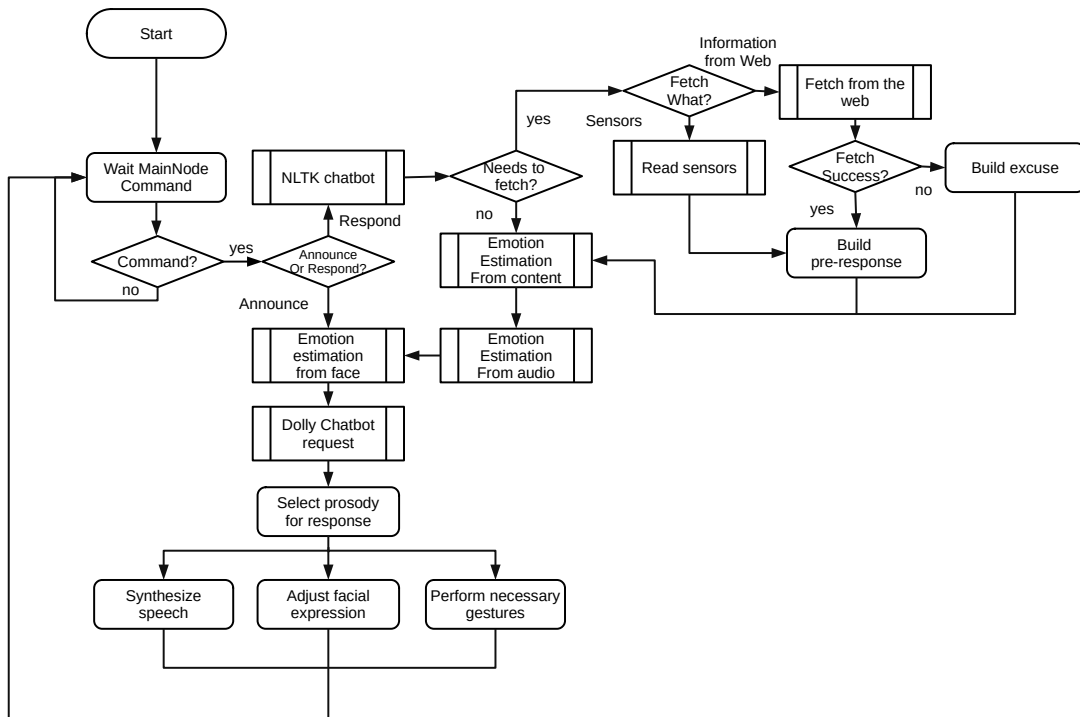


Fig. 8.22: Plantroid's Dialogue Manager.

question to the GPT-J Dolly model, it admits that it does not know and that the user should ask an expert or seek information from a reputable source. Misinformation can be very problematic and Plantroid was designed to avoid it at any cost, even if it might limit the performance of the system as a conversation partner.

MainNode also interacts with the CameraServer and performs sentiment analysis on the speech of users, trying to guess their internal emotional state. After understanding the emotional state of the interlocutor, Plantroid changes its facial expressions accordingly to the emotional state and the meaning it wants to convey; also selecting appropriate prosody parameters using GSIP as a way of improving the emotional state of listeners, if the current emotion is negative.

The proposed dialogue manager, thus, executes the algorithm displayed in Figure 8.22.

The many faces of Plantroid

Body language is very important in communication and, thus, if the social Plantroid was not able to display emotions through gestures and facial expressions, only voice could convey its

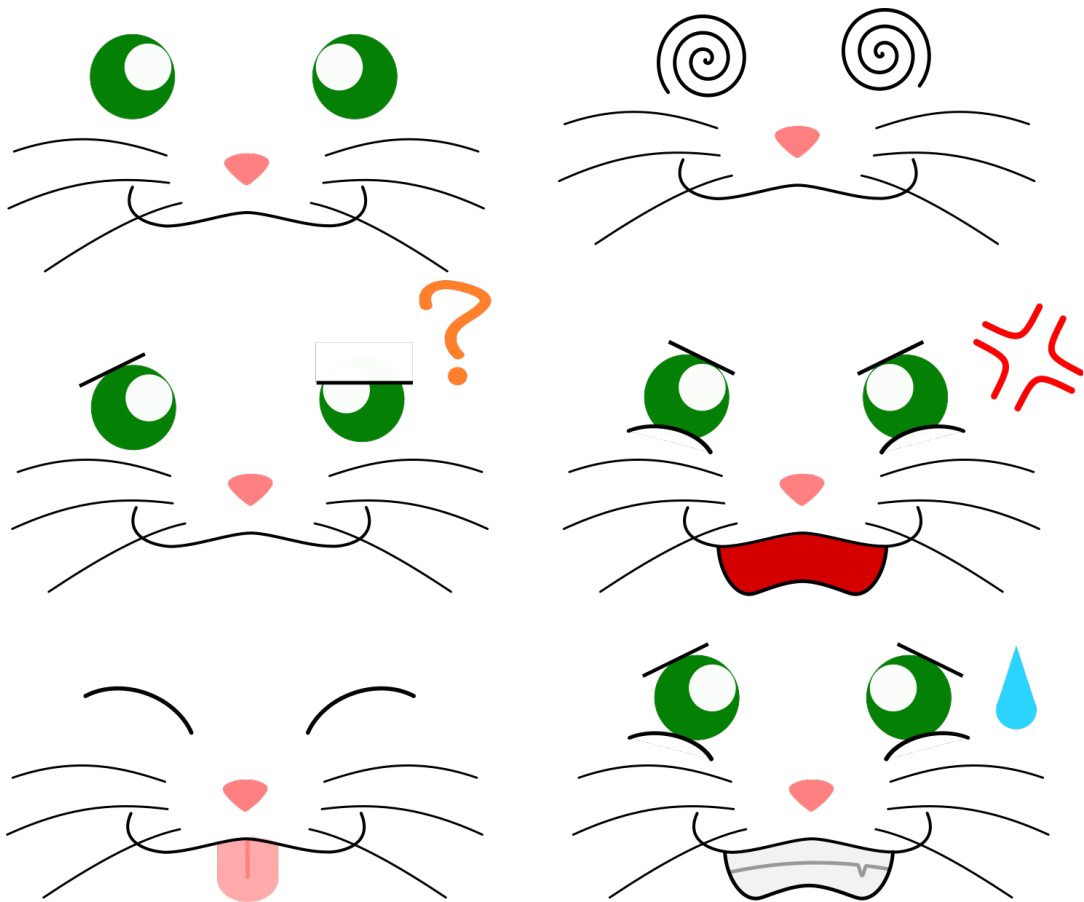


Fig. 8.23: Some of the possible facial expressions of Social Plantroid.

feelings and, thus, it could be perceived as insincere, if, for example, it says it is sorry with a neutral or happy facial expression. The facial expression and body language system is controlled by the Dialogue Manager, which is under the GUINode. Plantroid's face has a cartoonish cat-like appearance, but the system was developed in a way to easily allow to change the facial expressions. The face itself is modular, with right and left eyes, mouth, and emotional symbol, a character that emphasizes the current emotion displayed, such as an exclamation mark for surprise, question mark for confusion or doubt, among others. The system, albeit simple, is very powerful; there are 13 distinct eye types, 5 emotional symbols (6 with no symbol) and 4 mouths (5 with no mouth), allowing to generate 5,070 distinct facial expressions; a few of which can be seen in Figure 8.23.

第9章

Conclusion

This thesis investigated how speech characteristics and the appearance of embodied conversational agents shape human impression, with main focus on the study of how speech content (Gibberish or English), Phone choice and acoustic prosody characteristics immediately change the emotion of listeners. It also focused on how embodiment and anthropomorphism level of ECAs shape human immediate and post-hoc impression of the agents. Since previous research had glossed over the possible interference of preference for novelty affecting their results, I also investigated how embodiment and experience with robots affect human impression of three distinct representations of the same Social Plantroid entity. However, since it was necessary to develop a robot for the experiment and to test the developed prosody selection systems, this thesis also covered the development process of the Social Plantroid Robot and its end-to-end VGG-16-based navigation architecture.

9.1 Conclusions from “Talk to Kotaro”

In the “Talk to Kotaro” experiment, 37 participants from 10 different regions, speaking a total of 14 languages between them, contributed over 730 audio and video samples of their conversation with a 2D animated screen-based ECA, Kotaro. In order to investigate how gibberish speech whose phone distribution does not follow a traditional Yulean-like distribution and not a traditional syllabic structure, many different analysis were performed over the audiovisual data recorded in the “Talk to Kotaro” web-based crowdsourcing experiment. The research was mostly interested in the immediate emotional changes caused by listening to utterances $S(w, P)$ with distinct w vectors of IPA phones and the matrix of the associated acoustic prosody characteristics P , which were chosen accordingly to Algorithm 1. Moreover, we have also analyzed the average emotion displayed by volunteers while listening to Kotaro’s GS utterances, since it gives a very useful insight of how participants felt during the overall experiment, instead of focusing in their momentary emotional state. Moreover, the quantitative and qualitative investigation performed over the optional Likert scale experiment helped us understand and validate the results of the previous analyses.

By analyzing the facial expressions of volunteers in the video samples and the main features of their speech through the MFCC of the audio samples, we were able to verify the findings of [56] that gibberish speech is not very engaging for talking with conversational agents. The experiments yielded little to no positive emotional impact, as indicated by the negative average emotion

scores. While the diversity of valence responses suggests sporadic positive experiences, the prevailing estimated sentiment among participants was of impatience and frustration. The difficulty of changing the emotional state of participants and the mostly neutral and negative stance towards the prompts of the Likert scale questionnaire further reinforces the notion that engagement and overall experience provided by gibberish speech in a conversational setting were sub-optimal. The study's alignment with previous research underscores the challenge of forging engaging interactions with GS-speaking agents for adults, which suggests that it is not a recommended means of communication for a conversational setting, since conversations tend to feel one-sided, as highlighted by the attitude towards prompt P_6 .

Delving deeper into the analysis of prosody, interaction duration, and phone choice, attempts to understand what characteristics of the GS utterances generated the estimated impressions, the results of Subsections 5.3.2 and 5.5 show that the correlation between the prosodic parameters and the immediate emotion change on volunteers were not statistically relevant, barring a very weak correlation between pitch and arousal. Divergent impression patterns among participants from different cultural backgrounds (Japanese and Brazilian) suggest that the initial hypothesis of a cross-cultural preference for specific prosodic attributes does not hold, underscoring the complexity of cross-cultural communication preferences. However, further investigation on the effect of sample size on calculated p -values show that much more data are necessary to strengthen the finding of our work, which will be achieved through a longer user study.

The trained GRU_{phones} neural networks used for predicting valence and arousal changes from the tokenized IPA phones of the generated GS utterances achieved good performance for training and validation data; however, its generalization capabilities were lackluster. The interest in the pre-trained GRU_{phones} models lies in their learned embedding hyperspaces, particularly where each phone is positioned relative to other phones. The initial hypothesis to be tested in such analysis is that phones with close articulation location in the human mouth were expected to be close to each other in the valence and arousal embedding spaces, since it was expected that they would generate similar impressions on listeners. Yet, limitations stemming from the stochastic nature of the algorithm while selecting phones for Kotaro's utterances did not allow all phones to figure in the data set, and thus reduced the capacity to which deeper investigations can be performed. Nonetheless, through the K-means clustering method and by computing the distances among all

phones, we found out that the phones that are the closest to each other in the embedding space, more often than not, do not have a close articulation locus, showing no support for the original hypothesis. However, since not all phones were used in the experiment, Algorithm 1 needs to be modified to take into account phone frequency and which phones have not yet been used during interactions.

The results of Subsection 5.3.5 help to understand why most of the impressions presented by the research subjects were of low valence and low-to-moderate arousal, which is consistent with the mostly neutral or slightly negative attitude towards the ECA revealed by the analysis of the responses to the Likert scale prompts. It also shows that participants attribute a low level of intelligence to humanoid-like ECAs that only speak gibberish. The turn-based conversation was not a good interface for the research, since the participants did not particularly enjoy it, and for further experiments, VAD should be used to guarantee more natural conversations.

The route of predicting human impression from the gibberish speech patterns does not seem promising, as well as its use for human–computer and human–robot conversation, since research subjects seemed to not enjoy the experience, since they could not understand what the robot avatar was trying to convey. Even though research subjects were aware of the fact that the Kotaro avatar did not speak semantic speech, they seemed to still be trying to understand what meaning it was trying to convey. This way, it is necessary to compare the performance non-Yulean gibberish speech with Yulean gibberish speech and against other semantic-free utterances in order to better understand how it performs against other SFUs. Moreover, it is necessary to compare how those SFUs perform in a conversational vs. expressive setting, where conversational agents use SFUs to make the listeners believe the agent feels a certain way.

Regarding prosody selection, the *GSIP* system was not able to predict human impression very well for test data, showing a lack of generalization capabilities. Yet, it obtained a better performance than $MLP_{profile+prosody}$ and GRU_{phones} , showing that taking information about the conversation partner, acoustic prosody capabilities, and the phones of the GS utterance allows the system to make more accurate predictions, even though it shows a strong preference for small emotional change predictions, which is in accordance with most of the data set. It is necessary to take into account that the lack of correlation shown between prosody parameters and the clash between the attitudes towards prompts P_3 and P_6 for male respondents might hint towards a complicated re-

relationship between phone choice and impression. Another issue that is worth investigating is the validity of using facial expressions to estimate the emotional state of participants in low valence and arousal states, since it might be difficult to distinguish their actual emotion, since it is very close to neutrality.

9.2 Conclusions from the GSIP experiment

In order to address gaps in the first experiment and to test the developed GSIP system as a mean of generating adequate prosody for semantic-free and semantic speech, Phases 1 and 2 of the GSIP experiment were performed. In the experiment volunteers talked once again with Kotaro avatar, but had also to interact with two new agents, Plantroid Avatar and the Social Plantroid Robot, which spoke Gibberish Speech in P_1 and English Language during Phase P_2 . The prosody for said speech contents were select using three different methods, constant prosody, GSIP-based prosody selection and random prosody selection. In order to measure the performance of the prosody selection systems, gibberish speech itself and the agents themselves, two adapted Godspeed Scale Questionnaires were proposed together with ranking questionnaires and video recordings of the volunteers were taken, which allowed to estimate their emotional state from the displayed facial expressions. The experiment had 7 research hypotheses, 6 which were tested in phases P_1 and P_2 and the 7th one during phase P_3 , which was more of a very welcome bonus of investigating the role of the embodiment level of ECA, novelty bias and the perception of several characteristics about the agents.

From the obtained responses to the questionnaires and the video analysis, there is no support for H_1 – speech generated by the proposed system is perceived as more human-like than speech with constant prosody or with randomly generated prosody. It seems to be on par with prosody generated by other systems for Gibberish speech; it only seems less responsive, but such perception does not extend to semantic speech. However, regarding how natural the generated utterances are, GSIP performs worse than constant and random prosody selection in P_2 .

Regarding H_2 – speech generated by the proposed system generates more positive impression on volunteers than speech with constant or random prosody patterns; from the results of the video analysis, it is possible to say that for Gibberish speech, the GSIP-based prosody selection system

does not generate a better experience than constant prosody, despite generating more positive emotions than Random prosody. However, for semantic speech, GSIP seems to have a better average performance, but, by breaking the performance down to a per-user basis, results are identical to P_1 , which seems to show that more intense positive emotion peaks skew the results in favor of GSIP.

For Hypothesis H_3 , GSIP successfully generated the desired impression on volunteers, even outperforming the results obtained for training data. Since it is the first system of its kind, it has, by default, the best human impression prediction performance, but it allows for automatically verifying how much an utterance will change the emotional state of humans.

In order for H_4 – test subjects are more lenient with a non-humanoid looking avatar regarding semantic-free speech and eventual bad selection of prosody; to be true, volunteers would have had to be more lenient while evaluating the perceived intelligence, pleasantness and responsiveness of the more animal-like agents and the effect was quite the opposite. Volunteers evaluated the pet-like robots worse, showing that the intelligence evaluation is linked to the anthropomorphism degree of the agent.

Since the GSIP-based prosody selection system achieved similar performance for the virtual Plantroid Avatar and for the robot in P_1 and even outperformed the Avatar in P_2 , the data seems to support that the system can be successfully used in Physical robots, there is support for hypothesis H_5 .

Finally, regarding Hypothesis H_6 , it came as a surprise that, on average, GSIP outperformed constant and random prosody selection methods in several instances, while it failed to do so for gibberish speech. However, since H_6 presumed that H_n , $n = 1, 2, 3, 4$ would hold true, it is not possible to say that there is support for it. Nonetheless, it achieves better results than other systems on average, but not on a by-case basis. Such results suggest that, for semantic speech, some volunteers found the lack of prosody variation to be less desirable than some variety, given by the GSIP-based system; and both were better than the great and sudden prosody changes caused by the random prosody selection.

Such results, however, do not imply that GSIP does not generate good prosody, since its predictions of the impression of volunteers were accurate. It means, however that it is necessary to speed up the predictions of the neural networks in order to allow for generating a larger pool of candidate prosodies and, given that the result of the Talk to Kotaro experiment suggests good prosody se-

lection to be a personalization task, to develop an online learning version of the proposed neural network, capable of learning while interacting with users.

Moreover, the system must be retrained with all the new gathered data, which has higher quality than the data obtained through crowdsourcing, given every participant had different system setups and environment conditions. By including all data, performance is bound to improve.

Regarding Phase P_3 , analysis has shown that the ECAs were well rated by participants and, from the adapted Godspeed scale questionnaire, a strong negative correlation between the perception of friendliness and the level of physical embodiment for male volunteers with at least some experience with robots ($\tau_C = -0.38$, p -value=0.046) was obtained. For male volunteers with little experience with robots (level 1) a very strong negative correlation between embodiment level and perceived intelligence ($\tau_C = -0.88$, p value=0.053) was shown. A strong negative correlation was found between perceived intelligence of all ECA and level of experience with robots for male participants ($\tau_C = -0.38$, p value=0.002).

For female participants, there was a moderate correlation between the perceived responsiveness of all ECA and experience with robots, with $\tau_C = 0.24$ and p -value=0.042. Regarding the ranking, male respondents showed a negative correlation between the level of experience with robots and the ranking given to the screen agent, with τ_C of -0.47 and p -value=0.049; the correlation between gender and the ranking of the screen agent is quite strong, with $\tau_C = 0.75$ and p -value=0.083, i.e. male respondents tend to give the screen agent a higher ranking.

Regarding the ranking, male respondents showed a negative correlation between the level of experience with robots and the ranking given to the screen agent, with τ_C of -0.47 and p -value=0.049, showing preference for it. The holographic agent, contrary to expectation, was deemed to be the least favorite for every level of experience with robots. Both the robot and the holographic agents were novel for inexperienced volunteers; and the robot received better rankings from this group, suggesting that both novelty and physical embodiment play a role in shaping volunteers' perceptions. However, how much each aspect contributes still requires future investigation. Many volunteers attributed not ranking the holographic agent higher due to its small size. Thus, for future studies it is necessary to obtain a larger display or use an alternative medium for holography, such as VR or AR goggles.

The analysis performed on the interaction recordings show that the holographic display, on av-

erage, was more engaging than other agents for volunteers with higher levels of experience with robots (levels 2 and 3), and the robot was more engaging for those with little or no previous experience with robots, seemingly showing support for hypothesis $H_{7,a}$. However, the obtained correlations between the valence and the experience with robots were weak or moderate, while some strong correlations were obtained between the embodiment level and valence suggest otherwise. Moreover, no relevant correlation was found between arousal and embodiment level or arousal and experience with robots, which warrants further investigation.

Such results do not fully support hypothesis H_7 , since it only shows that the agents that volunteers are familiar with have lower questionnaire scores and rankings, but there is no increase in the score for the agents that are novel. For the inexperienced participants, both the robot and the holographic agents were novel; and the robot received better ratings and rankings from this group, showing, together with the significant correlation with embodiment level, that both familiarity with robots and physical embodiment play a role in shaping volunteers' perceptions. However, how much each aspect contributes requires future investigation.

The analysis performed on the recordings of the interactions with the agents showed average negative values for valence, showing that the users had serious, concentrated facial expressions when listening to the responses of the agents, but the variances were quite high, showing some moments where the mood was lighter. The immediate response to the interaction was quite different from that of the ranking questionnaire, with the holographic display having a better result for all groups except those who had no experience with robots, showing partial support for H_7 , but also showing that physical embodiment matters when both experiences are new.

However, since the holographic display was quite small, it is necessary either to obtain a larger display for future experiments or to use a smaller robot and reduce the size of the screen agent; however, the second solution does not seem ideal since it worsens the performance of the other agents.

9.3 Conclusions from the development of Plantroid

The Social Plantroid robot was successfully developed taking into account the findings of previous experiments and other research, in order to provide a test platform for the findings of this research and to test the findings of other researchers' works, helping to create a full embodied

conversation agent design philosophy where conversational agents are also workers and have a meaning outside of their owners. It was used in the GSIP experiment, receiving good scores in the adapted Godspeed Scale and ranking questionnaire, being the favorite agent of many volunteers – specially those who had no previous experience with robots. Although that is not captured by the questionnaires, many volunteers mentioned, after the experiment was finished, that the robot was considered to be cute; the exact goal of giving it a pet-like appearance. Multiple volunteers also stated to the robot itself that it looked like a cat, specially during phase P_1 .

The VGG-16-based architecture described in this work has shown that it can successfully learn how to avoid obstacles in the different environments using from monocular gray scale images from the data generated with the artificial potential field method. It has learned how to implicitly behave in a similar fashion of the artificial potential field method, that is, does not need to explicitly estimate the distance of obstacles and then run the APF method. The proposed architecture reduces the need for other distance measuring sensors such as LiDAR, sonars *etc.*

It has also shown to be resistant to changes in the position of obstacles in the environment and changes in the lighting conditions. The weaknesses to the proposed architecture is that, for some obstacles, it has shown an extreme avoidance, mainly for the tables of Σ_1 . Its greatest weakness, however, is inherent to vision-only methods - the robot can only avoid obstacles inside its field of view. For the present architecture, this can be mitigated by using cameras with a wider field of view or a neural network architecture with memory, such as LSTM, so it remembers obstacles that were inside the field of view of the robot, but exited it as the robot rotates or moves closer, which remains as a future work.

9.4 Future Research

Since the results from the “Talk to Kotaro” and the subsequent GSIP evaluation experiment strongly suggested that GS is not the best communication means for ECA, at least for adults, it becomes necessary then to test its performance in a passive setting, as was done in previous research, such as in works [100, 98, 97, 96]. Moreover, since there are no other works comparing other modalities of SFU to Gibberish Speech, it is necessary to obtain a better understanding on how their performance compare to each other, allowing researchers and engineers to choose the most appropriate SFU for a given application.

Regarding the selection of prosody, it is necessary to re-evaluate the prosody selection strategy being employed. For future experimentation, another possible route to establish rapport and show that the agent understands the emotion behind the words said by users is to use the prosody synchronicity approach [200], where the conversational agent establishes rapport by copying the prosodic parameters of the interlocutor.

The present experiments, have not investigate all the capabilities of the Plantroid robot, such as its dialogue management system, bodily language manager, soil monitoring and plant health prediction based of the recordings of the soil nutrients. Testing all such systems remains as very necessary future work, which will greatly contribute to the further adoption of the proposed robot platform. Its navigation systems, however, were tested in an objective manner, but the subjective performance necessary to evaluate the perception of users about comfort levels of staying in the same environment where the robot navigate, if the proxemics consideration of having the robot looking into the interlocutor eyes and keeping its position while talking, leaving to the human to adjust the distance between the two according to his or her liking, all remain to be tested and validated. Such experiments remain as future works, as well as comparing the proposed architecture against other end-to-end solutions, such as reinforcement-learning approaches, and to extend the present architecture to a recurrent one, enabling the system to remember encountered obstacles and avoid them after they are no longer visible to the onboard camera as the robot navigates through the environment.

Acknowledgments

I dedicate this thesis first and foremost to the Holy Family of Nazareth, Our Lord Jesus Christ, Holy Mary Mother of God, and Saint Joseph, Terror of Demons. Christ the King, Savior of Humanity, Lord of the Universe, countless titles sing Your endless glory; but they are not enough. Humbly, I offer this doctoral thesis to Your greater Glory. *Non nobis Domine, non nobis, sed nomini Tuo da Gloriam.* To Our Lady and Saint Joseph, I also dedicate this work; without your timeless intercession and example, I would never have had the strength to reach where I have.

To my parents, José and Elisiane, no word, gesture, or expression of human love would suffice to thank you for all the love, affection, and dedication directed at me throughout my life – my doctoral journey was no exception. My first teachers, heroes, and life examples, I always hope to make you proud with my choices and achievements. I love you with all my being. This thesis is dedicated to you.

I also dedicate this to my brother Lucas, for all the support, affection, love, and companionship during our lives. Without your unconditional friendship, without all our adventures and mutual support, this life would have little flavor. I would not have achieved this doctorate without you; this victory is also yours.

I dedicate this to the memory of my late grandmothers Maria and Elizabeth and my so beloved grandfather Dedé, whose unconditional love, sweetness, and endless giving colored my childhood with dreams and paved my way with gentle examples that guide me to this day. I hope that even from Heaven I can still make you proud. I can't wait for our reunion in Eternity.

I dedicate this to my grandfather, Antonio, for instilling in me a love of reading and knowledge; and to my grandfather Jose, for his example of hard work and dedication. I also dedicate this to the memory of Uncle José, our so dearly missed Uncle Zé, who always prayed for me and guided me during the time we were together in the Youth Teams of Our Lady.

I dedicate this to my aunts, Maricarmem, Antonio, Vera, Gentil, and Cristina for all the affection; as well as to my beloved cousins, Daniela, Felipe, Lilian, Andre, Tiago, and Bárbara for all the companionship during our lives. I also dedicate this to all other family members who supported me so much, encouraged me, and taught me; this victory is all of ours.

I want to thank all my teachers who taught me so much, each contributing a piece to the endless puzzle that is the corpus of human knowledge. Your examples and teachings were essential for my life and for this doctoral thesis. Particularly, I want to thank my advisor, Professor Dr. Ikuo Mizuuchi who, much more than an excellent teacher and advisor, was a true mentor and friend. I thank you from the heart for all the valuable lessons, both in Engineering and in Life.

I also want to thank Professor Dr. Paulo Rosa, from the Instituto Militar de Engenharia (English: Military Institute of Engineering) for introducing me to Professor Mizuuchi; this thesis would not even have started were it not for his priceless help; I cannot express my gratitude enough.. I also want to thank my friend and master Daniel Runkel, who more than guiding me in the world of

martial arts and physical exercise, gave great examples for life.

I want to thank my Japanese teachers, Monique, Nina Bassous, and Tomoko Hongo, without whose teachings my life in Japan would have been much more complicated, if not impossible.

I want to thank Professors Patricia McGahan and Yukiko Horikiri for all the support during the doctorate, advising, guiding, helping to get more participants for my experiments; it is no exaggeration to say that this thesis is due to you. However, more than that, I want to thank you for the enormous friendship and affection you had for me, certainly, I would not have reached the end of this course without you.

I want to thank Chieko san from the International House of TUAT, who took care of me in my first moments here in Japan; we exchanged many words, smiles, and sweets.

I want to thank my friends in Brazil, especially Gustavo Amaral, João Pedro, Marcos Paulo, Rafael Prallon, João Vitor, and Victoria Porto, with whom I lived so many fantastic adventures that I will carry with me for life.

I want to thank the members of the Team of Our Lady of Fatima of the EJNS, Amanda, Carol, Rafa, Daiana, Thays, Orfeu Lazaro, Aunt Ilma, Aunt Cris, Uncle Marcelo, and Deacon Moyses for all the love, affection, and prayers before and during the doctorate; this victory is ours by the grace of God! Together until Heaven!

I want to thank my friends from TUAT who were by my side in my happiest and saddest moments, always giving support and encouragement that were essential for the completion of this thesis; in particular Kevin Yapri, HaiLiang Liew, WingSum Lo, Juan Pablo, Miumi Maezawa, Riana Yachi, Carina, Mint, Yee Jingzu, Abdullah Adham, Salehuddin, Simeon Capy, Pablo Osorio, Iliia Radchenko, Jameil Magomnang, Kazuki Sekine, Tomohiro Shimazaki, Yudai Yamanoto, and Kazumi Kumagai; and so many others who supported me.

I also want to thank Fathers Kato and Vincenzo and Bishop Andrea Lembo for all the spiritual support and guidance during the doctorate, may God always bless and protect them.

I wanted to thank my great friends, Diego Lopes and Akira Hirata, from the Parish of Fuchu for becoming true brothers in such a short time of coexistence; thank you very much for the adventures, laughter, and support in difficult times.

I would also like to thank all the staff of the Tokyo University of Agriculture and Technology, the International House of Koganei, the Hitotsubashi University Dormitory, and the Keyaki Dormitory for all the essential support for the completion of this doctoral course.

I would like to thank the WISE Program and the FLOuRISH Fellowship for all the incredible learning opportunities, for introducing me to some of the most amazing people I have ever met and for financially supporting my research.

I thank ASEAN and the FLOuRISH Fellowship for providing the necessary scholarships for me to subsist during the doctoral course. Finally, I would like to thank the People of Japan for their warm welcome and for maintaining quality universities with generous scholarships through their collective efforts.

This work was supported by the Proposal-based Project budget of the Doctoral Program for World-leading Innovative & Smart Education of Tokyo university of Agriculture and Technology (WISE Program of TUAT): "Excellent Leader Development for Super Smart Society by New Industry Creation and Diversity" and by the Support for Pioneering Research Initiated by the Next Generation of FLOuRISH Institute, Tokyo University of Agriculture and Technology, both granted by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

関連図書

- [1] Atsushi Deguchi, et al. What is society 5.0. *Society*, Vol. 5, pp. 1–23, 2020.
- [2] Heiner Lasi, et al. Industry 4.0. *Business & information systems engineering*, Vol. 6, No. 4, pp. 239–242, 2014.
- [3] GGKWSIR Karunarathne, KADT Kulawansa, and MFM Firdhous. Wireless communication technologies in internet of things: a critical evaluation. In *2018 International conference on intelligent and innovative computing applications (ICONIC)*, pp. 1–5. IEEE, 2018.
- [4] James Lester, Karl Branting, and Bradford Mott. Conversational agents. In *The practical handbook of internet computing*, pp. 220–240, 2004.
- [5] Srinu Janarthanam. *Hands-on chatbots and conversational UI development: build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. Packt Publishing Ltd, 2017.
- [6] Selma Yilmazyildiz, Robin Read, Tony Belpeame, and Werner Verhelst. Review of semantic-free utterances in social human–robot interaction. *International Journal of Human-Computer Interaction*, Vol. 32, No. 1, pp. 63–85, 2016.
- [7] Markus Schwenk and Kai O Arras. R2-d2 reloaded: a flexible sound synthesis system for sonic human-robot interaction design. In *The 23rd IEEE international symposium on robot and human interactive communication*, pp. 161–167. IEEE, 2014.
- [8] RA Caroro, AB Garcia, and CS Namoco. A text-to-speech using rule-based and data-driven prosody techniques with concatenative synthesis of the philippines ’ bisaya dialect. *International Journal of Applied Engineering Research*, Vol. 10, No. 19, pp. 40209–40215, 2015.
- [9] Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yonghui Wu. Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6699–6703. IEEE, 2020.
- [10] Enrico Zovato, Alberto Pacchiotti, Silvia Quazza, and Stefano Sandri. Towards emotional speech synthesis: A rule based approach. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.
- [11] Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 853–864, 2022.

- [12] Selma Yilmazyildiz, et al. Emogib: emotional gibberish speech database for affective human-robot interaction. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011.
- [13] Antonio Galiza Cerdeira Gonzalez, WingSum Lo, and Ikuo Mizuuchi. Talk to kotaro: a web crowdsourcing study on the impact of phone and prosody choice for synthesized speech on human impression. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 244–251. IEEE, 2022.
- [14] Ikuo Mizuuchi, et al. Development of musculoskeletal humanoid kotaro. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006*. IEEE, 2006.
- [15] Vilayanur S. Ramachandran and Edward M. Hubbard. Synaesthesia—a window into perception, thought and language. *Journal of consciousness studies*, Vol. 8, No. 12, pp. 3–34, 2001.
- [16] Finley Lau, Deepak Gopinath, and Brenna D. Argall. A javascript framework for crowd-sourced human-robot interaction experiments: Remotehri. In *Association for the Advancement of Artificial Intelligence (2020)*, 2020.
- [17] Sonia Chernova, et al. Crowdsourcing human-robot interaction: Application from virtual to physical worlds. In *2011 RO-MAN*. IEEE, 2011.
- [18] Tetsunari Inamura and Yoshiaki Mizuchi. Sigverse: A cloud-based vr platform for research on multimodal human-robot interaction. *Frontiers in Robotics and AI*, Vol. 8, p. 158, 2021.
- [19] Hoang-Long Cao, et al. Dualkeepon: a human–robot interaction testbed to study linguistic features of speech. *Intelligent Service Robotics*, Vol. 12, No. 1, pp. 45–54, 2019.
- [20] Tom Ziemke. What ’ s that thing called embodiment? In *Proceedings of the annual meeting of the cognitive science society*, 第 25 卷, 2003.
- [21] Kerstin Fischer, Katrin S. Lohan, and Kilian Foth. Levels of embodiment: Linguistic analyses of factors influencing hri. *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012.
- [22] S. Ter Stal, L. L. Kramer, M. Tabak, H. op den Akker, and H. Hermens. Design features of embodied conversational agents in ehealth: a literature review. *International Journal of Human-Computer Studies*, Vol. 138, p. 102409, 2020.
- [23] Nicole Salomons, et al. The impact of an in-home co-located robotic coach in helping people make fewer exercise mistakes. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022.

- [24] Kerstin Fischer, Katrin S. Lohan, and Kilian Foth. Levels of embodiment: Linguistic analyses of factors influencing hri. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012.
- [25] Dianne C. Berry, Laurie T. Butler, and Fiorella de Rosis. Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, Vol. 63, No. 3, pp. 304–327, 2005.
- [26] Masato Yuasa, Satoshi Nishiki, and Ikuo Mizuuchi. 1a1-q05 自律移動可能な果樹栽培型 plantroid の開発 (農業用ロボット・メカトロニクス). ロボティクス・メカトロニクス講演会講演概要集 2013, pp. _1A1-Q05_1. 一般社団法人 日本機械学会, 2013.
- [27] Masato Yuasa and Ikuo Mizuuchi. A control method for a swarm of plant pot robots that uses artificial potential fields for effective utilization of sunlight. *Journal of Robotics and Mechatronics*, Vol. 26, No. 4, pp. 505–512, 2014.
- [28] S Santos Valle and Josef Kienzle. *Agriculture 4.0—Agricultural robotics and automated equipment for sustainable crop production*. FAO: Food and Agriculture Organization of the United Nations, 2020.
- [29] Veronique Beckers, L. Poelmans, A. Van Rompaey, and N. Dendoncker. The impact of urbanization on agricultural dynamics: a case study in belgium. *Journal of Land Use Science*, Vol. 15, No. 5, pp. 626–643, 2020.
- [30] Peter BR Hazell. Urbanization, agriculture, and smallholder farming. In *Agriculture & Food Systems to 2050: Global Trends, Challenges and Opportunities*, pp. 137–160. World Scientific, 2019.
- [31] Massimiliano Agovino, Mariaconcetta Casaccia, Mariateresa Ciommi, Maria Ferrara, and Katia Marchesano. Agriculture, climate change and sustainability: The case of eu-28. *Ecological Indicators*, Vol. 105, pp. 525–543, 2019.
- [32] Onyekachukwu Akaeze and Dilip Nandwani. Urban agriculture in asia to meet the food production challenges of urbanization: A review. *Urban Agriculture & Regional Food Systems*, Vol. 5, No. 1, p. e20002, 2020.
- [33] Gustavo Belforte, R Deboli, Paolo Gay, Pietro Piccarolo, and D Ricauda Aimonino. Robot design and testing for greenhouse applications. *Biosystems Engineering*, Vol. 95, No. 3, pp. 309–321, 2006.
- [34] Robert McDougall, Paul Kristiansen, and Romina Rader. Small-scale urban agriculture results in high yields but requires judicious management of inputs to achieve sustainability. *Proceedings of the National Academy of Sciences*, Vol. 116, No. 1, pp. 129–134, 2019.

- [35] Siv Lene Gangenes Skar, Rocío Pineda-Martos, Axel Timpe, Bernd Pölling, Katrin Bohn, Mart Külvik, Cecília Delgado, CMG Pedras, TA Paço, Mirjana Čujić, et al. Urban agriculture as a keystone contribution towards securing sustainable and healthy development for cities in the future. *Blue-Green Systems*, Vol. 2, No. 1, pp. 1–27, 2020.
- [36] Ritesh Kumar Singh, Mohammad Hasan Rahmani, Maarten Weyn, and Rafael Berkvens. Joint communication and sensing: A proof of concept and datasets for greenhouse monitoring using lorawan. *Sensors*, Vol. 22, No. 4, p. 1326, 2022.
- [37] Masato Yuasa and Ikuo Mizuuchi. A control method for a swarm of plant pot robots that uses artificial potential fields for effective utilization of sunlight. *Journal of Robotics and Mechatronics*, Vol. 26, No. 4, pp. 505–512, 2014.
- [38] Robert Sparrow and Mark Howard. Robots in agriculture: prospects, impacts, ethics, and policy. *precision agriculture*, Vol. 22, No. 3, pp. 818–833, 2021.
- [39] Stephen J DeCanio. Robots and humans—complements or substitutes? *Journal of Macroeconomics*, Vol. 49, pp. 280–291, 2016.
- [40] Paul Baxter, Grzegorz Cielniak, Marc Hanheide, and Pål From. Safe human-robot interaction in agriculture. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 59–60, 2018.
- [41] Zhihao Shen, Armagan Elibol, and Nak Young Chong. Understanding nonverbal communication cues of human personality traits in human-robot interaction. *IEEE/CAA Journal of Automatica Sinica*, Vol. 7, No. 6, pp. 1465–1477, 2020.
- [42] PAULO G PINHEIRO, JOSUE JUNIOR GUIMARAES RAMOS, and GUSTAVO HENRIQUE DE OLIVEIRA. Smile and talk-ana, the architecture of a robot that verbally and nonverbally understands you. In *Congresso Brasileiro de Automática-CBA*, Vol. 1, 2019.
- [43] Eva Blessing Onyeulo and Vaibhav Gandhi. What makes a social robot good at interacting with humans? *Information*, Vol. 11, No. 1, p. 43, 2020.
- [44] OMRON. B5t human vision components (hvc-p2) (accessed: 2022/08/24), 2022.
- [45] AdaFruit. Adafruit mlx90640 ir thermal camera - overview (accessed: 2022/08/24), 2022.
- [46] Christoph Bartneck. Why do all social robots fail in the market?, 2020.
- [47] Yuri DV Yasuda, Luiz Eduardo G Martins, and Fabio AM Cappabianco. Autonomous visual navigation for mobile robots: A systematic literature review. *ACM Computing Surveys (CSUR)*, Vol. 53, No. 1, pp. 1–34, 2020.

- [48] Shuzhi Sam Ge and Yun J Cui. Dynamic motion planning for mobile robots using potential field method. *Autonomous robots*, Vol. 13, No. 3, pp. 207–222, 2002.
- [49] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pp. 2722–2730, 2015.
- [50] Stefan A. Dumitru, Luige Vladareanu, Tian Hong Yan, and Chen Kun Qi. Mobile robot navigation techniques using potential field method in unknown environments. In *Monitoring, Controlling and Architecture of Cyber Physical Systems*, Vol. 656 of *Applied Mechanics and Materials*, pp. 388–394. Trans Tech Publications Ltd, 11 2014.
- [51] Carlos Morais, Tiago Pereira do Nascimento, Alisson V Brito, and Gabriel Basso. A 3d anti-collision system based on artificial potential field method for a mobile robot. In *ICAART (1)*, pp. 308–313, 2017.
- [52] HY Lee, Hann Woei Ho, and Ye Zhou. Deep learning-based monocular obstacle avoidance for unmanned aerial vehicle navigation in tree plantations: Faster region-based convolutional neural network approach. *Journal of Intelligent & Robotic Systems*, Vol. 101, pp. 1–18, 2021.
- [53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [54] Eun-Sook Jee, et al. Sound design for emotion and intention expression of socially interactive robots. *Intelligent Service Robotics*, Vol. 3, No. 3, pp. 199–206, 2010.
- [55] Selma Yilmazyildiz, et al. Review of semantic-free utterances in social human–robot interaction. *International Journal of Human-Computer Interaction*, Vol. 32, No. 1, pp. 63–85, 2016.
- [56] Azumi Ueno, Kotaro Hayashi, and Ikuo Mizuuchi. Impression change on nonverbal non-humanoid robot by interaction with humanoid robot. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019.
- [57] Matthew P. Aylett, Yolanda Vazquez-Alvarez, and Skaiste Butkute. Creating robot personality: effects of mixing speech and semantic free utterances. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020.
- [58] David Bell, Tania Koulouri, Salvatore Lauria, Robert D. Macredie, and Jonathan Sutton. Microblogging as a mechanism for human–robot interaction. *Knowledge-Based Systems*, Vol. 69, pp. 64–77, 2014.

- [59] Nathan Tsoi, et al. Challenges deploying robots during a pandemic: An effort to fight social isolation among children. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021.
- [60] Mariah L. Schrum, et al. Four years in review: Statistical practices of likert scales in human-robot interaction studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020.
- [61] C. Vinola and K. Vimaladevi. A survey on human emotion recognition approaches, databases and applications. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, Vol. 14, No. 2, pp. 24–44, 2015.
- [62] Yafei Sun, et al. Authentic emotion detection in real-time video. In *International Workshop on Computer Vision in Human-Computer Interaction*. Springer, 2004.
- [63] Huadong Li and Hua Xu. Deep reinforcement learning for robust emotional classification in facial expression recognition. *Knowledge-Based Systems*, Vol. 204, p. 106172, 2020.
- [64] Isha Talegaonkar, et al. Real time facial expression recognition using deep learning. In *Proceedings of International Conference on Communication and Information Processing (ICCIP)*, 2019.
- [65] International Phonetic Association and International Phonetic Association Staff. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.
- [66] Jonathan Duddington and R. Dunn. espeak text to speech. Web publication: <http://espeak.sourceforge.net>, 2012.
- [67] Hiroya Fujisaki. Prosody, models, and spontaneous speech. *Computing prosody: Computational models for processing spontaneous speech*, pp. 27–42, 1997.
- [68] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, Vol. 39, No. 6, p. 1161, 1980.
- [69] Paul Ekman. Are there basic emotions? *Psychological review*, Vol. 99, No. 3, pp. 550–553, 1992.
- [70] Abhijit Mondal and Swapna S Gokhale. Mining emotions on plutchik’s wheel. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 1–6. IEEE, 2020.
- [71] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction

- in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, Vol. 127, No. 6-7, pp. 907–929, 2019.
- [72] Jason S Haukoos and Roger J Lewis. Advanced statistics: bootstrapping confidence intervals for statistics with “ difficult ” distributions. *Academic emergency medicine*, Vol. 12, No. 4, pp. 360–365, 2005.
- [73] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.
- [74] Harald Steck and Tommi Jaakkola. Bias-corrected bootstrap and model uncertainty. *Advances in neural information processing systems*, Vol. 16, , 2003.
- [75] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, Vol. 82, No. 397, pp. 171–185, 1987.
- [76] Thomas Diccio and Bradley Efron. More accurate confidence intervals in exponential families. *Biometrika*, Vol. 79, No. 2, pp. 231–245, 1992.
- [77] Jordan Zlatev. Levels of meaning, embodiment, and communication. *Cybernetics & Human Knowing*, Vol. 16, No. 3-4, pp. 149–174, 2009.
- [78] Margaret Wilson. Six views of embodied cognition. *Psychonomic bulletin & review*, Vol. 9, pp. 625–636, 2002.
- [79] Robert Dale. GPT-3: What ’ s it good for? *Natural Language Engineering*, Vol. 27, No. 1, pp. 113–118, 2021.
- [80] Hsin-I Liao, Su-Ling Yeh, and Shinsuke Shimojo. Novelty vs. familiarity principles in preference decisions: Task-context of past experience matters. *Frontiers in Psychology*, Vol. 2, p. 43, 2011.
- [81] John R Evans. Improving photosynthesis. *Plant physiology*, Vol. 162, No. 4, pp. 1780–1793, 2013.
- [82] Xia Hao, Jingdun Jia, Jiaqi Mi, Si Yang, Abdul Mateen Khattak, Lihua Zheng, Wanlin Gao, and Minjuan Wang. An optimization model of light intensity and nitrogen concentration coupled with yield and quality. *Plant Growth Regulation*, Vol. 92, No. 2, pp. 319–331, 2020.
- [83] Sonal Mathur, Divya Agrawal, and Anjana Jajoo. Photosynthesis: response to high temperature stress. *Journal of Photochemistry and Photobiology B: Biology*, Vol. 137, pp. 116–126, 2014.

- [84] ON Sherstneva, VA Vodeneev, LM Surova, EM Novikova, and VS Sukhov. Application of a mathematical model of variation potential for analysis of its influence on photosynthesis in higher plants. *Biochemistry (Moscow) Supplement Series A: Membrane and Cell Biology*, Vol. 10, No. 4, pp. 269–277, 2016.
- [85] An Long, Jiang Zhang, Lin-Tong Yang, Xin Ye, Ning-Wei Lai, Ling-Ling Tan, Dan Lin, and Li-Song Chen. Effects of low pH on photosynthesis, related physiological parameters, and nutrient profiles of citrus. *Frontiers in plant science*, Vol. 8, p. 185, 2017.
- [86] Stefan D. Kalev and Gurpal S. Toor. Chapter 3.9 - the composition of soils and sediments. In Béla Török and Timothy Dransfield, editors, *Green Chemistry*, pp. 339–357. Elsevier, 2018.
- [87] Allen R Overman and Richard V Scholtz III. *Mathematical models of crop growth and yield*. CRC Press, 2002.
- [88] Magnus Egerstedt, Xiaoming Hu, and Alexander Stotsky. Control of mobile platforms using a virtual vehicle approach. *IEEE transactions on automatic control*, Vol. 46, No. 11, pp. 1777–1782, 2001.
- [89] Keith Brown. *Concise encyclopedia of pragmatics*. Elsevier, 2009.
- [90] Niels Ole Bernsen, Hans Dybkjær, and Laila Dybkjær. Cooperativity in human-machine and human-human spoken dialogue. *Discourse processes*, Vol. 21, No. 2, pp. 213–236, 1996.
- [91] Keith Brown. *Concise encyclopedia of pragmatics*. Elsevier, 2009.
- [92] S. M. Bhagya P. Samarakoon, M. A. Viraj J. Muthugala, and A. G. Buddhika P. Jayasekara. A review on human-robot proxemics. *Electronics*, Vol. 11, No. 16, 2022.
- [93] Jonathan Mumm and Bilge Mutlu. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, pp. 331–338, 2011.
- [94] Michael Argyle and Janet Dean. Eye-contact, distance and affiliation. *Sociometry*, pp. 289–304, 1965.
- [95] Michael L Walters, Kerstin Dautenhahn, Kheng Lee Koay, Christina Kaouri, R te Boekhorst, Chrystopher Nehaniv, Iain Werry, and David Lee. Close encounters: Spatial distances between people and a robot of mechanistic appearance. In *5th IEEE-RAS International Conference on Humanoid Robots, 2005.*, pp. 450–455. IEEE, 2005.

- [96] Selma Yilmazyildiz, David Henderickx, Bram Vanderborght, Werner Verhelst, Eric Soetens, and Dirk Lefebber. Multi-modal emotion expression for affective human-robot interaction. In *Proceedings of the Workshop on Affective Social Speech Signals (WASSS 2013), Grenoble, France, 2013*.
- [97] Selma Yilmazyildiz, Lukas Latacz, Wesley Mattheyses, and Werner Verhelst. Expressive gibberish speech synthesis for affective human-computer interaction. In *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings 13*, pp. 584–590. Springer, 2010.
- [98] Selma Yilmazyildiz, Georgios Athanasopoulos, Georgios Patsis, Weiyi Wang, Meshia Cédric Oveneke, Lukas Latacz, Werner Verhelst, Hichem Sahli, David Henderickx, Bram Vanderborght, et al. Voice modification for wizard-of-oz experiments in robot-child interaction. In *Proceedings of the workshop on affective social speech signals, Grenoble, 2013*.
- [99] Yuri Tambovtsev and Colin Martindale. Phoneme frequencies follow a yule distribution. *SKASE Journal of Theoretical Linguistics*, Vol. 4, No. 2, pp. 1–11, 2007.
- [100] Weiyi Wang, Georgios Athanasopoulos, Selma Yilmazyildiz, Georgios Patsis, Valentin Enescu, Hichem Sahli, Werner Verhelst, Antoine Hiolle, Matthew Lewis, and Lola Canamero. Natural emotion elicitation for emotion modeling in child-robot interactions. In *WOCCI*, pp. 51–56, 2014.
- [101] Ganapreeta Renunathan Naidu, Syaheerah Lebai Lutfi, Amal Abdulrahman Azazi, Jaime Lorenzo-Trueba, and Juan Manuel Montero Martinez. Cross-cultural perception of spanish synthetic expressive voices among asians. *Applied Sciences*, Vol. 8, No. 3, 2018.
- [102] Fabrice Malfrere, Thierry Dutoit, and Piet Mertens. Automatic prosody generation using suprasegmental unit selection. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.
- [103] Joram Meron. Prosodic unit selection using an imitation speech database. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.
- [104] Tuomo Raitio, Ramya Rasipuram, and Dan Castellani. Controllable neural text-to-speech synthesis using intuitive prosodic features. *arXiv preprint arXiv:2009.06775*, 2020.
- [105] Mireille Fares. Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 743–747, 2020.
- [106] Max Morrison, Zeyu Jin, Justin Salamon, Nicholas J Bryan, and Gautham J Mysore. Controllable neural prosody synthesis. *arXiv preprint arXiv:2008.03388*, 2020.

- [107] Yuanhao Yi, Lei He, Shifeng Pan, Xi Wang, and Yujia Xiao. Prosodyspeech: Towards advanced prosody model for neural text-to-speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7582–7586. IEEE, 2022.
- [108] Younggun Lee, Azam Rabiee, and Soo-Young Lee. Emotional end-to-end neural speech synthesizer, 2017.
- [109] Jianhua Tao and Aijun Li. Emotional speech generation by using statistic prosody conversion methods. *Affective Information Processing*, pp. 127–141, 2009.
- [110] Se-Yun Um, Sangshin Oh, Kyunguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. Emotional speech synthesis with rich and granularized control. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7254–7258. IEEE, 2020.
- [111] Parthana Sarma and Shovan Barma. Review on stimuli presentation for affect analysis based on eeg. *IEEE Access*, Vol. 8, pp. 51991–52009, 2020.
- [112] Stavros G Vougioukas. Agricultural robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, Vol. 2, No. 1, pp. 365–392, 2019.
- [113] Jawad Iqbal, Rui Xu, Hunter Halloran, and Changying Li. Development of a multi-purpose autonomous differential drive mobile robot for plant phenotyping and soil sensing. *Electronics*, Vol. 9, No. 9, p. 1550, 2020.
- [114] Ahmed Hassan, Rao M Asif, Ateeq Ur Rehman, Zuhaib Nishtar, Mohammed KA Kaabar, and Khan Afsar. Design and development of an irrigation mobile robot. *IAES International Journal of Robotics and Automation*, Vol. 10, No. 2, p. 75, 2021.
- [115] Sk Bilal, V Meghanethra, and Sk Fareed. Swarm robot for irrigation based application. *International Journal of Engineering Applied Sciences and Technology*, Vol. 5, No. 6, pp. 331–334, 2020.
- [116] AO Adeodu, OP Bodunde, IA Daniyan, OO Omitola, JO Akinyoola, and UC Adie. Development of an autonomous mobile plant irrigation robot for semi structured environment. *Procedia Manufacturing*, Vol. 35, pp. 9–15, 2019.
- [117] Anu Kumari and H. Prasanna Kumar. Agrobot - the healthy farming. *International Journal of Research in Engineering, Science and Management*, Vol. 2, pp. 260–263, 2019.
- [118] OP Bodunde, UC Adie, OM Ikumapayi, JO Akinyoola, and AA Aderoba. Architectural design and performance evaluation of a zigbee technology based adaptive sprinkler irrigation robot. *Computers and Electronics in Agriculture*, Vol. 160, pp. 168–178, 2019.

- [119] Hema Nagaraja, Reema Aswani, and Monisha Malik. Plant watering autonomous mobile robot. *IAES International Journal of Robotics and Automation*, Vol. 1, No. 3, p. 152, 2012.
- [120] Ilker Ünal, Önder Kabaş, and Salih Sözer. Real-time electrical resistivity measurement and mapping platform of the soils with an autonomous robot for precision farming applications. *Sensors*, Vol. 20, No. 1, p. 251, 2020.
- [121] Mehdi Hussain, Syed Hassan Abbas Naqvi, Salman Hassan Khan, and Muhammad Farhan. An intelligent autonomous robotic system for precision farming. In *2020 3rd International Conference on Intelligent Autonomous Systems (ICoIAS)*, pp. 133–139, 2020.
- [122] Sankarananth S and Arun R S. A smart cable-driven parallel robot assistant for individual plant care in farming. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, pp. 295–301, 2022.
- [123] Goran Kitić, Damir Krklješ, Marko Panić, Csaba Petes, Slobodan Birgermajer, and Vladimir Crnojević. Agrobot lala—an autonomous robotic system for real-time, in-field soil sampling, and analysis of nitrates. *Sensors*, Vol. 22, No. 11, p. 4207, 2022.
- [124] Shiva Gorjian, Hossein Ebadi, Max Trommsdorff, H Sharon, Matthias Demant, and Stephan Schindele. The advent of modern solar-powered electric agricultural machinery: A solution for sustainable farm operations. *Journal of cleaner production*, Vol. 292, p. 126030, 2021.
- [125] B R Jerosheja and C Mythili. Solar powered automated multi-tasking agricultural robot. In *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pp. 1–5, 2020.
- [126] Giuseppe Quaglia, Carmen Visconte, Leonardo Sabatino Scimmi, Matteo Melchiorre, Paride Cavallone, and Stefano Pastorelli. Design of a ugv powered by solar energy for precision agriculture. *Robotics*, Vol. 9, No. 1, p. 13, 2020.
- [127] G. M. Sharif Ullah Al-Mamun, Md Imran Hossain, Md.Rokib Hasan, Ashfaqur Rahman, Sokhorio MargonD ' Costa, Al Jubair Hossain, Md.Rabiul Islam, Md.Tusher Alam, Arpita Hoque. Performance analysis of multipurpose agrobot. In *2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, pp. 1–4, 2019.
- [128] Ayumi Kawakami, Koji Tsukada, Keisuke Kambara, and Itiro Siio. Potpet: Pet-like flow-erpot robot. In *Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '11, p. 263–264, New York, NY, USA, 2010. Association for Computing Machinery.
- [129] Juan Pablo Vasconez, Leonardo Guevara, and Fernando Auat Cheein. Social robot navigation based on hri non-verbal communication: A case study on avocado harvesting. In

- Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, p. 957–960, New York, NY, USA, 2019. Association for Computing Machinery.
- [130] Juan P Vasconez, George A Kantor, and Fernando A Auat Cheein. Human–robot interaction in agriculture: A survey and current challenges. *Biosystems engineering*, Vol. 179, pp. 35–48, 2019.
- [131] Paul Baxter, Grzegorz Cielniak, Marc Hanheide, and Pål From. Safe human-robot interaction in agriculture. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, p. 59–60, New York, NY, USA, 2018. Association for Computing Machinery.
- [132] Thomas B Sheridan. A review of recent research in social robotics. *Current opinion in psychology*, Vol. 36, pp. 7–12, 2020.
- [133] Iolanda Leite, Carlos Martinho, and Ana Paiva. Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, Vol. 5, No. 2, pp. 291–308, 2013.
- [134] AIST. PARO Therapeutic Robot — parorobots.com. <http://www.parorobots.com/>, 2014. [Accessed 26-Aug-2022].
- [135] SoftBank Robotics. NAO the humanoid and programmable robot | SoftBank Robotics — softbankrobotics.com. <https://www.softbankrobotics.com/emea/en/nao>, 2018. [Accessed 26-Aug-2022].
- [136] SoftBank Robotics. Pepper the humanoid and programmable robot | SoftBank Robotics — softbankrobotics.com. <https://www.softbankrobotics.com/emea/en/pepper>, 2018. [Accessed 26-Aug-2022].
- [137] Cesar Vandevelde, Jelle Saldien, Maria-Cristina Ciocci, and Bram Vanderborght. Ono, a diy open source platform for social robotics. In *International conference on tangible, embedded and embodied interaction*, 2014.
- [138] Jelle Saldien, Stan Notebaert, Cesar Vandevelde, and Dries Bovijn. Demonstration of oporo’s grid system: Design a working social robot in only 2 hours. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, p. 46–47, New York, NY, USA, 2017. Association for Computing Machinery.
- [139] Victor C. Dibia, Maryam Ashoori, Aaron Cox, and Justin D. Weisz. Tjbot: An open source diy cardboard robot for programming cognitive systems. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, p. 381–384, New York, NY, USA, 2017. Association for Computing Machinery.

- [140] Micol Spitale, Chris Birmingham, R. Michael Swan, and Maja J Matarić. Composing harmoni: An open-source tool for human and robot modular open interaction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3322–3329, 2021.
- [141] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics*, Vol. 7, No. 3, pp. 347–360, 2015.
- [142] Diego Casas-Bocanegra, Daniel Gomez-Vargas, Maria J. Pinto-Bernal, Juan Maldonado, Marcela Munera, Adriana Villa-Moreno, Martin F. Stoelen, Tony Belpaeme, and Carlos A. Cifuentes. An open-source social robot based on compliant soft robotics for therapy with children with asd. *Actuators*, Vol. 9, No. 3, 2020.
- [143] Daniel Hayosh, Xiao Liu, and Kiju Lee. Woody: Low-cost, open-source humanoid torso robot. In *2020 17th International Conference on Ubiquitous Robots (UR)*, pp. 247–252, 2020.
- [144] Michael Ferguson, Nick Webb, and Tomek Strzalkowski. Nelson: a low-cost social robot for research and education. In *Proceedings of the 42nd ACM technical symposium on Computer science education*, pp. 225–230, 2011.
- [145] Diego Corrochano, Enzo Ferrari, María Antonia López-Luengo, and Vanessa Ortega-Quevedo. Educational gardens and climate change education: An analysis of spanish pre-service teachers’ perceptions. *Education Sciences*, Vol. 12, No. 4, 2022.
- [146] Aubrey Shick. Romibo robot project: An open-source effort to develop a low-cost sensory adaptable robot for special needs therapy and education. In *ACM SIGGRAPH 2013 Studio Talks, SIGGRAPH ’13*, New York, NY, USA, 2013. Association for Computing Machinery.
- [147] Michael Suguitan and Guy Hoffman. Blossom: A handcrafted open-source robot. *J. Hum.-Robot Interact.*, Vol. 8, No. 1, mar 2019.
- [148] Selma Yilmazyildiz, Robin Read, Tony Belpaeme, and Werner Verhelst. Review of semantic-free utterances in social human–robot interaction. *International Journal of Human-Computer Interaction*, Vol. 32, No. 1, pp. 63–85, 2016.
- [149] Furi Sawaki, Kentaro Yasu, and Masahiko Inami. Flona: Development of an interface that implements lifelike behaviors to a plant. In *International Conference on Advances in Computer Entertainment Technology*, pp. 557–560. Springer, 2012.
- [150] Janine Stocker, Aline Veillat, Stéphane Magnenat, Francis Colas, and Roland Siegwart. Towards adaptive robotic green plants. In *Conference Towards Autonomous Robotic Systems*, pp. 422–423. Springer, 2011.

- [151] Varun Tolani, Somil Bansal, Aleksandra Faust, and Claire Tomlin. Visual navigation among humans with optimal control as a supervisor. *IEEE Robotics and Automation Letters*, Vol. 6, No. 2, pp. 2288–2295, 2021.
- [152] Lingyan Ran, Yanning Zhang, Qilin Zhang, and Tao Yang. Convolutional neural network-based robot navigation using uncalibrated spherical images. *Sensors*, Vol. 17, No. 6, p. 1341, 2017.
- [153] T Ran, L Yuan, and JB Zhang. Scene perception based visual navigation of mobile robot in indoor environment. *ISA transactions*, Vol. 109, pp. 389–400, 2021.
- [154] Ye-Hoon Kim, Jun-Ik Jang, and Sojung Yun. End-to-end deep learning for autonomous navigation of mobile robot. In *2018 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–6. IEEE, 2018.
- [155] Daniel Dugas, Olov Andersson, Roland Siegwart, and Jen Jen Chung. Navdreams: Towards camera-only rl navigation among humans, 2022.
- [156] Long Ma and Yanqing Zhang. Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015.
- [157] Miikka Silfverberg, Lingshuang Jack Mao, and Mans Hulden. Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, 2018.
- [158] Prudhvi Yenigalla, Anurag Kumar, Shubham Tripathi, Chandan Singh, Supratik Kar, and Jithendra Vepa. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In *Interspeech*, pp. 3688–3692, September 2018.
- [159] Zhuo Chen, Mayank Jain, Yuzong Wang, Mike L. Seltzer, and Christopher Fuegen. Joint Grapheme and Phoneme Embeddings for Contextual End-to-End ASR. In *INTERSPEECH*, pp. 3490–3494, 2019.
- [160] International Phonetic Association. Ipa charts and sub-charts in four fonts. Web publication: https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/IPA_charts_E.html, 2021. retrieved on March 8th.
- [161] Jayden L. Macklin-Cordes and Erich R. Round. Re-evaluating phoneme frequencies. *Frontiers in psychology*, p. 3181, 2020.
- [162] Matt Diamond. Recorderjs, 2016.
- [163] Dimitrios Kollias and Stefanos Zafeiriou. A multi-component cnn-rnn approach for dimensional emotion recognition in-the-wild. In *arXiv preprint arXiv:1805.01452*, 2018.

- [164] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, Vol. PP, No. 99, pp. 1–1, 2017.
- [165] Susan R. Fussell, Sara Kiesler, Leslie D. Setlock, and Victoria Yew. How people anthropomorphize robots. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, HRI '08*, p. 145–152, New York, NY, USA, 2008. Association for Computing Machinery.
- [166] Leila Takayama. Making sense of agentic objects and teleoperation: In-the-moment and reflective perspectives. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, HRI '09*, p. 239–240, New York, NY, USA, 2009. Association for Computing Machinery.
- [167] M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (tess). Borealis, 2020. DRAFT VERSION.
- [168] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, Vol. 13, No. 5, p. e0196391, 2018.
- [169] Shafiq Haq and Philip J.B. Jackson. Multimodal emotion recognition. In Wenwu Wang, editor, *Machine Audition: Principles, Algorithms and Systems*, chapter 17, pp. 398–423. IGI Global, 2010.
- [170] Kenneth J Berry, Janis E Johnston, Sammy Zahran, and Paul W Mielke. Stuart ’ s tau measure of effect size for ordinal variables: Some methodological considerations. *Behavior research methods*, Vol. 41, pp. 1144–1148, 2009.
- [171] Antonio Galiza Cerdeira Gonzalez, Wing-Sum Lo, and Ikuo Mizuuchi. The impression of phones and prosody choice in the gibberish speech of the virtual embodied conversational agent kotaro. *Applied Sciences*, Vol. 13, No. 18, p. 10143, 2023.
- [172] Mahalanobis Prasanta Chandra, et al. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, Vol. 2, pp. 49–55, 1936.
- [173] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pp. 747–748. IEEE, 2020.
- [174] Irwin Pollack, James M Pickett, and William H Sumby. On the identification of speakers by voice. *the Journal of the Acoustical Society of America*, Vol. 26, No. 3, pp. 403–406, 1954.

- [175] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, Vol. 106, No. 26, pp. 10587–10592, 2009.
- [176] Gordon Briggs. *Overselling: Is appearance or behavior more problematic*, 2015.
- [177] Cody Canning, Thomas J Donahue, and Matthias Scheutz. Investigating human perceptions of robot capabilities in remote human-robot team tasks based on first-person robot video feeds. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4354–4361. IEEE, 2014.
- [178] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, Vol. 1, No. 1, pp. 161–187, 2017.
- [179] Deborah A Cobb-Clark and Stefanie Schurer. The stability of big-five personality traits. *Economics Letters*, Vol. 115, No. 1, pp. 11–15, 2012.
- [180] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. 2008.
- [181] Michael L Walters, Dag S Syrdal, Kerstin Dautenhahn, René Te Boekhorst, and Kheng Lee Koay. Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, Vol. 24, pp. 159–178, 2008.
- [182] Michael Laakasuo, Jussi Palomäki, and Nils Köbis. Moral uncanny valley: A robot ’ s appearance moderates how its decisions are judged. *International Journal of Social Robotics*, Vol. 13, No. 7, pp. 1679–1688, 2021.
- [183] Astrid Weiss and Christoph Bartneck. Meta analysis of the usage of the godspeed questionnaire series. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 381–388. IEEE, 2015.
- [184] Aleksandra Kalinowska, Patrick M Pilarski, and Todd D Murphey. Embodied communication: How robots and people communicate through physical interaction. *Annual Review of Control, Robotics, and Autonomous Systems*, Vol. 6, pp. 205–232, 2023.
- [185] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. Embodiment and human-robot interaction: A task-based perspective. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 872–877. IEEE, 2007.

- [186] Christoph Bartneck, Juliane Reichenbach, and van A Breemen. In your face, robot! the influence of a character 's embodiment on how users perceive its emotional expressions. In *Design and Emotion 2004*, 2004.
- [187] Jakub Zlotowski and Christoph Bartneck. The inversion effect in hri: Are robots perceived more like humans or objects? In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 365–372. IEEE, 2013.
- [188] John G Daugman. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on acoustics, speech, and signal processing*, Vol. 36, No. 7, pp. 1169–1179, 1988.
- [189] Volodymyr Agafonkin. Polylabel: a fast algorithm for finding the pole of inaccessibility of a polygon. Please cite this software using these metadata.
- [190] Daniel Garcia-Castellanos and Umberto Lombardo. Poles of inaccessibility: A calculation algorithm for the remotest places on earth. *Scottish Geographical Journal*, Vol. 123, No. 3, pp. 227–233, 2007.
- [191] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [192] T Eiter and H Mannila. Computing discrete frechet distance. Technical report, Tech. Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, 1994.
- [193] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- [194] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. OpenAI, 2018.
- [195] Alain Vázquez, Asier López Zorrilla, Javier Mikel Olaso, and María Inés Torres. Dialogue management and language generation for a robust conversational virtual coach: Validation and user study. *Sensors*, Vol. 23, No. 3, 2023.
- [196] Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell, and Matei Zaharia. Hello dolly: Democratizing the magic of chatgpt with open models. *Databricks Blog*, March 2023. Accessed on May 31, 2023.
- [197] Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Databricks Blog*, April 2023. Accessed on May 31, 2023.

-
- [198] Minhyeok Lee. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, Vol. 11, No. 10, p. 2320, 2023.
- [199] Gernot Beutel, Eline Geerits, and Jan T Kielstein. Artificial hallucination: Gpt on lsd? *Critical Care*, Vol. 27, No. 1, p. 148, 2023.
- [200] Shogo Nishimura, Takuya Nakamura, Wataru Sato, Masayuki Kanbara, Yuichiro Fujimoto, Hirokazu Kato, and Norihiro Hagita. Vocal synchrony of robots boosts positive affective empathy. *Applied Sciences*, Vol. 11, No. 6, 2021.