UNIVERSITY OF CALIFORNIA
Santa Barbara

ROVER The Reactive Observant Vacuous Emotive Robot

A Project submitted in partial satisfaction
of the requirements for the degree of

Masters in Science

in

Media Arts and Technology

by

Hannah Wolfe

Committee in Charge:

Marko Peljhan, Co-Chair

JoAnn Kuchera-Morin, Co-Chair

Matthew Turk

July 2016

The Project of
Hannah Wolfe is approved:

Matthew Turk

JoAnn Kuchera-Morin, Committee Co-Chairperson

Marko Peljhan, Committee Co-Chairperson

July 2016

ROVER The Reactive Observant Vacuous Emotive Robot

by

Hannah Wolfe

To Lady Catterley for the hours spent sitting

on my lap purring while I wrote this.

Acknowledgements

I would like to thank Professor JoAnn Kuchera-Morin, of the Media Arts and Technology Program at University of California, Santa Barbara. Prof. Kuchera-Morin was always incredibly supportive and available, meeting with her weekly kept me on track and focused.

I would also like to thank Professor Marko Peljhan, of the Media Arts and Technology Program at University of California, Santa Barbara. Prof. Peljhan was helpful in consolidating ideas and his detailed feedback on my thesis was incredibly constructive. The project could not have been completed nor the pilot study done without Systemics Lab's financial support.

I would also like to Professor Matthew Turk for being an invaluable member of my committee, pushing me in a more technical direction, Professor Tobias Hollerer and Professor Yon Visell for help defining, running, and interpreting the results of the pilot study, Dr. Matthew Wright for pushing me to look at sound and emotion at a deeper level, and Professor Marcos Novak and Professor George Legrady for their feedback on structural and visualization design.

Thanks to Dr. Michael Mangus for being a bottomless resource of information, and a different perspective on my research. Thanks to Kenny Kim and the 2013 WORLDCHANGING class for technical help with the initial stages of the project. Also, thanks to Sahar Sajadieh for helping me focus my research.

Finally, I'd like to thank Pamela Wolfe for copy editing my document, Robert Wolfe for being a sounding board for understanding conceptual ideas, and both for being my parents, as well as Skyler Kasko and Lady Catterley for emotional support.

Curriculum Vitæ

Hannah Wolfe

Education

2016	Master's of Science in Media Arts and Technology, University of California, Santa Barbara (Expected).
2009	Bachelors in Liberal Arts, Concentration in Visual Arts, Bennington College, Vermont.

Professional Experience

2014 – 2016	Research Assistant, University of California, Santa Barbara.
2013 – 2014	Teaching Assistant, University of California, Santa Barbara.
2010 – 2012	Programmer/Analyst, SunEdison, San Francisco CA.

Awards

2014 & 2015	Mosher Foundation Fellowship, AlloSphere Research Group.
2012	Dean's Fellowship, University of California, Santa Barbara.

Selected Publications

- Rosli, M. H. W., Yerkes, K., Wright, M., Wood, T., Wolfe, H., Roberts, C.,& Estrada, F. R. (2015). Ensemble feedback instruments. In Proceedings of the International Conference on New Interfaces for Musical Expression.

Abstract

ROVER The Reactive Observant Vacuous Emotive Robot

Hannah Wolfe

ROVER, the Reactive Observant Vacuous Emotive Robot, is an art installation that explores mobile embodied interaction through expressive sound. It can also be used to collect video data from users so we can learn about human and robot interaction, particularly emotive response to sound. While natural conversations and emotions are difficult to express through robots, their expression makes robots more relatable and predictable. This project explores emotive responses elicited by non-linguistic sound used as a universal language. ROVER was shown as a series of art installations/performances and was used in a pilot study in a non-art context. A system was defined for creating emotive sound based on research in music and linguistics, manipulating aspects of fundamental frequency, amplitude, timbre, and motive. The pilot study investigated this, focusing on the effect of mobile embodied interaction on emotive expressive responses to algorithmically generated non-linguistic utterances. It was found that "happy" sounds increased self-reported valence and "sad" sounds decreased it. Also, it was found that singing computers are comparatively more emotionally arousing than singing robots. This finding could be due to a variety of factors including volume, distance, input method, or position. Whether the robot moved did not affect arousal. These are promising results and present us with a precedent to continue with a full study focused on analysing the video collected by ROVER.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

To make the transition from industry to society, robots need to be able to interact with novice users. While natural conversations and emotions are difficult to express through robots, they make robots more relatable and predictable. (Scheutz et al. 2007) I believe that in order to navigate uncontrolled environments with untrained users who may or may not even speak the robot's language, it is important to find a universal language so that robots can communicate. Looking at non-linguistic communication as a universal language can aid in the understanding of what specific emotive responses are caused by non-linguistic sound. Using non-linguistic sounds that convey emotional meaning, like warning or calming sounds, will streamline communication between robots and people. Movement and mobility allow the robot to seek out interaction, instead of waiting to be approached.

There are two major classes of robots: industrial robots and service robots. While industrial robots have been around since 1961 (Bekey & Yuh 2008), personal robots first appeared in the early 1980s. Their main function was to teach people how robots work. (Bell 1985) The first service robot that had a job was the HelpMate service robot in 1988. (Evans et al. 1989) Personal helper robots are nowhere near the point where they can interact with people in non-controlled environments. (Royakkers & van Est 2015) Care robots for the home, along with social service robots in museums and restaurants have begun to appear since the mid-80s. (Pieskä et al. 2012) Since the breakthrough in human computer interaction in 1972 with the Xerox Alto, a research specific computer, and the Apple Lisa in 1983, the general populous has started interacting with a computer outside of the keyboard. (Wadlow 1981) This computer integrated aspects of Douglas C. Engelbart's on-Line System's mouse and GUI, refined at Xerox PARC. (Engelbart & English 1968)

The importance of affording the system the capability of interpreting human intention is essential for seamless interaction between the person and the machine. Robots, like computers, are no longer something that only trained professionals use to solve complex problems. It has been shown that people treat computers like humans, so computers need to be able to respond like humans. (Reeves & Nass 1996) If technology follows human social expectations, people will find the interaction enjoyable, and empowering. (Reeves & Nass 1996) People prefer to interact multimodally, so multimodal interfaces need to be provided. (Oviatt 1997)

This project begins the exploration of emotive reactive systems through the design of ROVER, the Reactive Observant Vacuous Emotive Robot, an interactive sculpture and experimental platform for human-robot interaction. This art installation is an exploration of the effect of active embodied interaction on emotive expressive responses to algorithmically generated non-linguistic units. This artwork can also be used to collect data from users to learn about human and robot interaction. Though the robot was constructed primarily for art installations and performances, a pilot study was performed to explore the effect of active embodied interaction on emotive expressive responses to algorithmically generated, non-linguistic utterances. The information gathered from the system was compared to related research in music and linguistics. The sounds are generated algorithmically so it is a self contained system with no need for storage, that is able to create an infinite number of sounds. The emotive response to aspects of audio is being tested, not the response to the melodies themselves. (Further discussed in 3.4.1) In particular, the emotive response to non-linguistic utterances when the participant approached the robot, when the robot approached the participant, and when the participant was at a computer were compared. We hypothesized that, through the empathy and engagement embodiment creates in participants, they would have a stronger emotional response to an embodied robot.

1.2 Motivation

People react 10 times quicker to sound than to visual cues, (Horowitz 2012*a*) and one of the fastest triggers for emotion is sound. (Horowitz 2012*b*) Visual communication either via body language or physical appearance has been extensively explored in research (Breazeal 2003) and creates many layers of complexity, particularly through construction of the robot when implementing the design. Multimodal interaction is "interaction with the virtual and physical environment through natural modes of communication". (Bourguet 2003) Multimodal interaction is needed because it allows for universal access. (Oviatt et al. 2004) Non-verbal communication was specifically not investigated because that would obfuscate the purpose of this study, which is sound and embodiment. Physical attributes were not investigated because in this project they are not dynamic and therefore cannot portray emotion. Since visual communication was not a part of the study, the physical structure did not need to change dependent on the emotions being tested. This allowed for quick prototyping and for the physical body of the robot to be purely constrained to functionality. This system for emotive response is designed to be integrated later with visual communication to create multimodal interaction.

The way that a robot moves affects a person's emotive response. (Saerbeck & Bartneck 2010) Feature cues and movement are the two main pieces of information that children use to define the difference between an inanimate and an animate object. (Opfer & Gelman 2010) Since movement would create a stronger sense of embodiment, it is expected that participants would relate to a mobile robot more readily than to a non-mobile robot. Without interaction with its environment, a robot is not a robot; it is just an object. This interaction could be in any form: audible, textual, visual, or physical. For each level of interaction (face vs non-face, movement vs stationary), a stronger sense of embodiment would be created. The concept of entitativity, where people subconsciously determine whether or not an entity is part of a group based on certain perceptual cues, was coined by Campbell in 1958 (Campbell 1958). This was expanded by Ip to physical similarity and goal/behavior similarity cues (Ip et al. 2006). Robots create entitativity through behavior cues such as emotive sound and movement.

Robots are now also moving out of industry where they work with trained engineers, and into more domestic settings, like the home, healthcare and entertainment. (Pieskä et al. 2012) Asimov's Laws describe the three laws of robotics which allow robots to interact with humans. The primary law is to keep humans safe and secondarily to obey humans and finally to keep itself safe. (Asimov 1950) In order to fit within the first and second laws, robots need to be easily usable, trainable, accessible and pleasant. For technology to be accessible, information needs to be conveyed in a variety of forms including visual, aural, and tactile. While there is a lot of research on how to create robots that emote through visual communication and representation, emoting vocally is an important part of how animals (humans included) interact with one another. (Darwin et al. 1998) One of the earliest references to this in science is in Darwin's "The Expression of the Emotions in Man and Animals" where he states that "With many kinds of animals, man included, the vocal organs are efficient in the highest degree as a means of expression." (Darwin et al. 1998)

The uncanny valley (see Figure 1.1) is an issue in all forms of emotive communication. (Mori et al. 2012) The quality of synthesized speech is worse than synthesized facial expressions. (Bartneck 2002) While trying to line up words with expressive audio factors, the voice typically falls into the uncanny valley. Therefore it is important to look for a way to convey emotion and create an emotional response in the viewer purely through non-linguistic auditory cues, paralanguage, and prosody. To create the most comprehensive way to communicate and generate

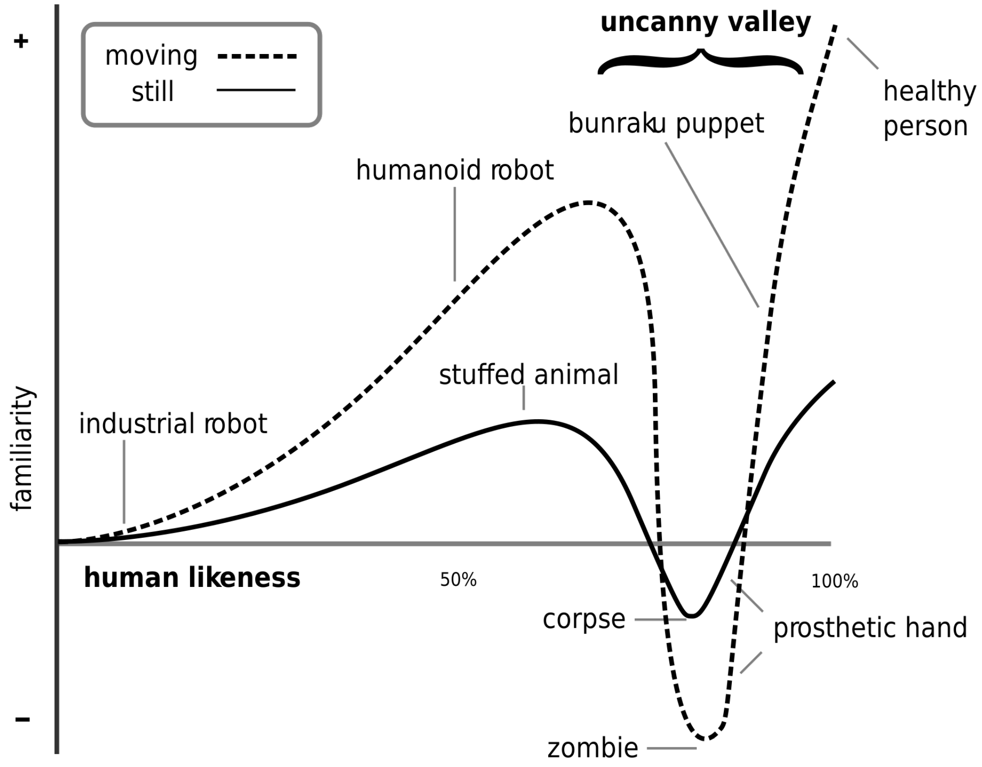


Figure 1.1: The uncanny valley, an aesthetic argument about the discomfort caused to the viewer when something seems mostly but not quite human. "Mori Uncanny Valley" by Smurrayinchester CC BY 3.0.

expressive responses, research was conducted in music, linguistics and psychology. (See Chapter 2.4) Using this information two different emotive expressive sounds were created to provoke responses from participants.

The project was initially inspired by a 2nd floor hallway of the California NanoSystems Institute building (Elings Hall) designed by architect Robert Venturi at the University of California, Santa Barbara. (ARCHIGUIDE 2000-2016) The space is poorly lit, desolate, antiseptic, and windowless. The project was a way to bring warmth and joy into the space. The first iteration of ROVER detected heat and moved toward people, looking for warmth and attempting to make them happy through song. (see Chapter 3) This version of ROVER was dog-like and was named after the traditional dog name. ROVER learned through operant conditioning, an incredibly effective way to modify dog behavior. B. F. Skinner described using operant conditioning to train pigeons to turn around when they saw the word turn and peck at the word peck. (Skinner 1951) Shi researched the best ways for a robot to initiate conversations in natural settings.(Shi et al. 2015) While the installation was originally designed for a gallery or a more natural setting, the study was not conducted in a natural setting. Because maintaining personal space is important in human robot interaction (Bethel & Murphy 2006), ROVER was designed to maintain a 3-4 foot distance from participants.

While we are creating a culture which consumes technology, it is also important to understand it. With a well designed product, the user should not need to know how it works in order to use it, however, products should be designed so that they can be explored, mastered and understood, via open sharing accessibility, if the user wishes to do so. The best way to understand something is to be able to hack it and make it oneself. Amateur radio enthusiasts are an early example of hacker culture, which led crystal control of radio transmitters to be a common and well controlled practice. (Brown 1996) The computer hobbyist movement in the 70s was where many of the early microcomputer founders started.(Levy 2001) Today's open hardware/software community emphasizes the use of technology to teach technology and make it accessible. In this project accessible technology was chosen so that we can give back to this community. All of the technology used is designed as a learning tool and can be ordered online or laser cut from acrylic, costing in total approximately 600 dollars. As part of this project it is being documented online so others can reproduce and modify this tool for studying human robot interaction. This will allow others to build from the platform and use it for other experiments or projects.

1.3 Goals of the System

The goal of ROVER is for it to be a robot that is responsive, observant, and emotive. The robot should respond in a period of time that is reasonable and respond when a person is present. The robot should be able to see people and detect people under varying conditions. The robot should also be a viable method to collect data from users. The robot should be able to express emotion and create an emotive expressive response in the viewer. ROVER should be an autonomous robot, so the robot needs a reasonable battery life. ROVER also needs to be able to map the space accurately and navigate the space without damaging itself or others.

ROVER is designed as an interactive art installation and has been presented at the Media Arts and Technology End of Year show in 2013, 2014, 2015 and 2016. It was also used for a pilot study to explore the effect of active embodied interaction on emotive expressive responses to algorithmically generated non-linguistic utterances.

The current iteration of ROVER is an interactive sculpture that navigates the space looking for people and playing them algorithmically generated music to learn what sounds provoke which emotive facial expressions, focusing on elation and sadness. Based on audio and emotion research, ROVER modulates audio qualities like timbre, fundamental frequency, contour, mode, and tempo. It also learns the space, mapping obstacles and finding people, while navigating using bump, proximity and heat sensors. (see Chapter 3) ROVER is built from an iRobot CREATE, a Raspberry Pi, a Raspberry Pi camera board, a Melexis 90620 (thermophile), an Arduino, proximity sensors and speakers.

Chapter 2

Related Work

2.1 Robots and Speech

Survey results have shown people prefer to communicate with robots via voice and they prefer that the voice be human-like. (Dautenhahn et al. 2005) (Khan 1998) (Ray et al. 2008) In film, robots with emotional vocal response typically fall into one of two categories, voice actors reading the lines with a filter (e.g. C3PO) or nonverbal computer generated sounds (e.g. R2D2).(Read 2014) Both Roehling and Xingyan propose systems for producing emotional natural language speech by robot but they have not been tested. (Roehling et al. 2006) (Li et al. 2009) A learning robot (see Figure 2.1) will receive more and better training data if it expresses emotion through statements and voice pre-recorded by the author. (Leyzberg et al. 2011) Niculescu et al. found that a robot dressed as a woman with a higher pitch voice was rated as having a more attractive voice

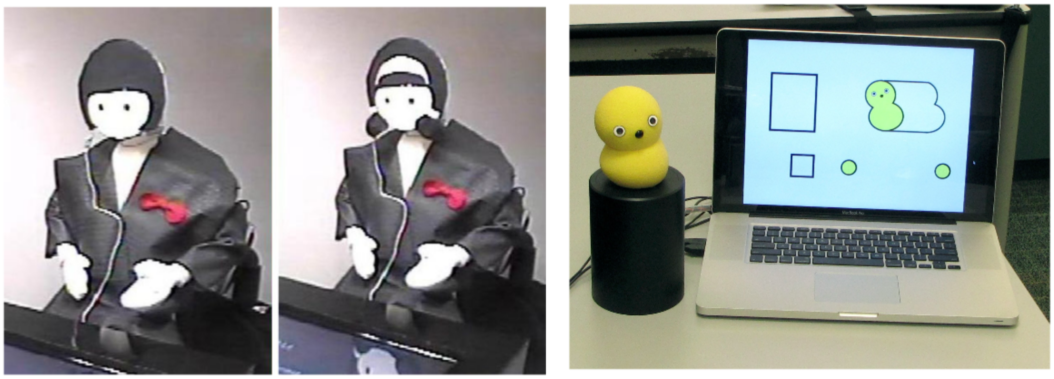


Figure 2.1: Niculescu et al. found that a robot dressed as a woman (pictured in the left two images) with a higher pitch voice was rated as having a more attractive voice when compared to one with a lower pitch voice. (Niculescu et al. 2013) A learning robot (pictured in the right image) will receive more and better training data if it expresses emotion through statements and voice pre-recorded by the author. (Leyzberg et al. 2011)

when compared to one with a lower pitch voice. (see Figure 2.1) It was also seen to be more aesthetically appealing, more outgoing, and having better social skills. (Niculescu et al. 2013) In designing a robot, having a voice actor record pre-generated speech can only cover a finite set of scenarios, however examples of generative systems have not been tested. I believe that a computer generated human voice that conveys emotion will fall into the uncanny valley of speech. There is no seminal paper on the uncanny valley of sound, but it is discussed in relation to creating horror in video games. (Grimshaw 2009)

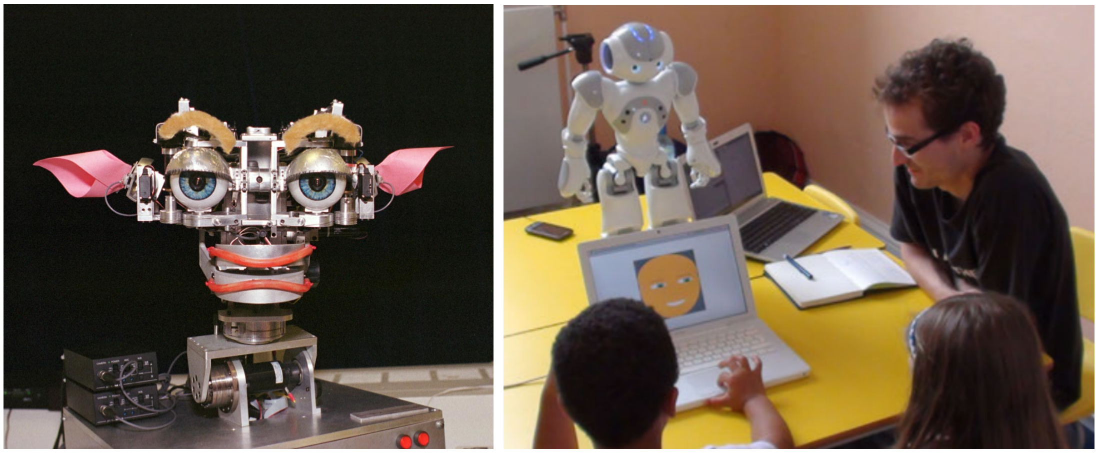


Figure 2.2: Cynthia Breazeal's robot Kismet (right) uses child-like utterances to reinforce emotions.(Breazeal 2004) Read has done extensive research on Non-linguistic utterances, with a Nao robot. (left) (Read 2014)

2.1.1 Gibberish as Non-verbal Communication

To simplify the problem and avoid the uncanny valley of speech, research has been done in creating gibberish and non-linguistic utterances that convey emotion. Cynthia Breazeal's robot Kismet (see Figure 2.2) uses child-like utterances to reinforce emotions. (Breazeal 2004) This system was created using DECtalk, a closed source product, so the algorithm is also closed source. Child-like babble was used convey emotions by Oudeyer that could be interpreted by people from different countries. (Pierre-Yves 2003) Oudeyer used the MBROLA synthesiser, a free product, to produce 30 sounds expressing Happiness, Sadness, Anger, Comfort and Calmness, using different input strings. The participants overall had a high accuracy in categorizing the sounds, though they confused the sounds representing comfort and calmness. Yilmazyildiz et al. proposed a way of creating gibberish using a database of sounds and a prosody template, but this algorithm was not tested. (Yilmazyildiz et al. 2006) In 2010 they created a different approach replacing the vowels in emotive sentences with different vowel sounds in a text to speech synthesizer. (Yilmazyildiz et al. 2010) They found that it is important to match the input language with the language of the text to speech synthesizer, and that it is easier to determine whether the statement is positive or negative if in the correct semantic context Yilmazyildiz combined these two methods in 2011 and found high recognition rates of the 7 emotions tested. (Yilmazyildiz et al. 2011) In his study in 2013 he found that recognition rates increase when facial expressions are included.(Yilmazyildiz et al. 2013) Voice actors, and text to speech synthesizers are heavily relied on for creating emotive gibberish.

2.1.2 Non-Linguistic Utterances

Read has done extensive research on non-linguistic utterances, with a Nao robot. (see Figure 2.2) Non-linguistic utterances are inexpensive computationally,(Read & Belpaeme 2012) and are cross-cultural. (Pierre-Yves 2003) Children will assign emotional meaning to non-linguistic utterances, though they sometimes disagree on the emotion (Read & Belpaeme 2012), while adults can categorize non-linguistic utterances.(Read & Belpaeme 2013) Non-linguistic utterances can create the appearance of a stronger emotional reaction in robots when in response to an action.(Read 2014) For robots that need to communicate emotions but not complex information, non-linguistic computer generated sounds are a simple and direct way to approach the problem. Earcons are nonverbal sounds that are used to convey information. Earcons have already been explored to convey weather information (Hermann et al. 2003) or to warn drivers(Larsson 2010). Even the phonemes of our words convey information. For example, when asked to associate words with different shapes, “maluma” was seen as round and “takete” as sharp by children who speak different languages. (Köhler 1929) These sounds can be algorithmically generated allowing for adaptation to different situations, but there are very few examples of algorithmically generated sounds that convey emotional meaning.

Sparky, a robot used by created by Mark Scheeff at Inverval Research Corp to investigate human robot interaction, made chirps but they were found to be confusing. (see Figure 2.3) Sparky is a small (50 cm tall) robot with an expressive face, a movable head on a long neck, and wheels. (Scheeff et al. 2002) Bartneck researched computer avatars conveying emotion through emotive utterances, but found that using visual cues or audio/visual cues was more effective. (Bartneck 2000) Toys like Keepon and WowWee’s line of robots use small databases of simple sounds for emotive expression. Keepon uses them as a means of attracting atten-

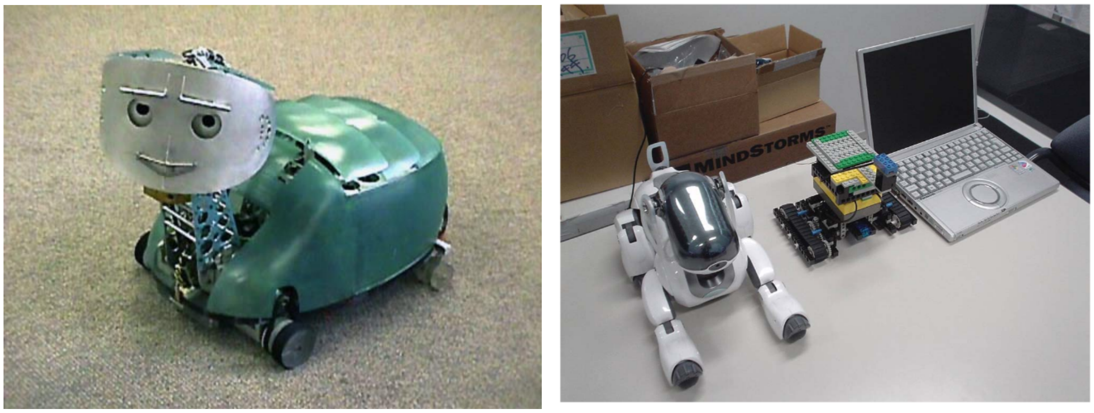


Figure 2.3: Sparky, (left) a robot used by created by Mark Scheeff at Inverval Research Corp to investigate human robot interaction, made chirps but they were found to be confusing. (Scheeff et al. 2002) Komatsu et al did an experiment where participants were asked to select the correct attitudes based on sounds produced by a MindStorms Robot, AIBO robot and laptop PC. (right) (Komatsu & Yamada 2011)

tion and as a response to sensory input. (Kozima et al. 2009) WowWee's robots use them as reactive behaviors to sensory input and to commands from a remote control. (Read 2014) These toys' expressions are not algorithmically generated and are no different from having a set of pre-recorded voice acted sounds.

Jee et al composed non-linguistic utterances based on music by modifying tempo, key, pitch, melody, harmony, and rhythm to represent happiness, sadness, fear and dislike. Results showed that composed music was very good at expressing emotion and worked best when paired with visual cues. (Jee et al. 2007) Jee et al later proposed an algorithmically generated musical system modifying tempo, pitch and volume to express joy, distress, shyness, irritation, expectation, dislike, pride and anger. Their system was not tested. (Jee et al. 2009) Jee et al later analyzed sounds from R2D2 and Wall-E and found that intonation, pitch and timbre were used to express emotions. They created 5 emotional expressions from this research and found that 55% of people felt that the sounds displayed intention and 80% felt the sounds displayed emotional expression. (Jee et al. 2010) These were not algorithmically generated and there was only one sound for each emotion or intention.

Komatsu et al did an experiment where participants were asked to select the correct attitudes based on sounds produced by a MindStorms Robot, AIBO robot and laptop PC. (see Figure 2.3) The results showed that the participants were better able to interpret PC sounds than sounds from the robots. (Komatsu & Yamada 2011) They also ran a study where a robot made a suggestion and followed it by a descending noise or flat noise. They found that participants were less likely to follow the suggestion with the descending noise. (Komatsu et al. 2010) They also found that people prefer earcons/non-linguistic utterances over language or paralanguage to display confidence. (Komatsu 2005)

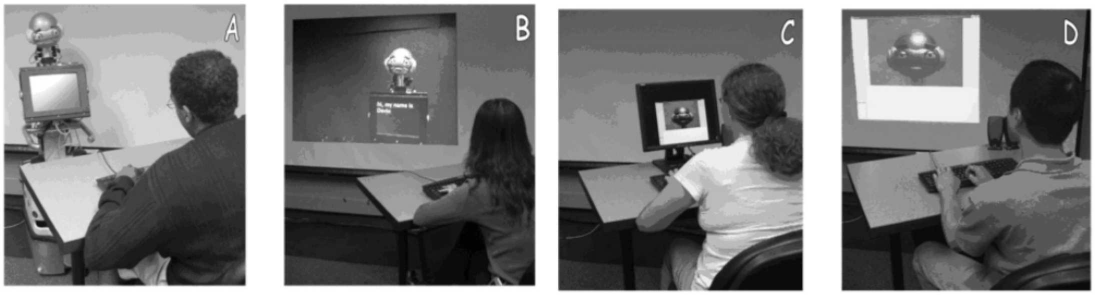


Figure 2.4: Participants felt a greater sense of presence, felt it was more lifelike, and disclosed less private information with an embodied robot versus an avatar. (Kiesler et al. 2008)

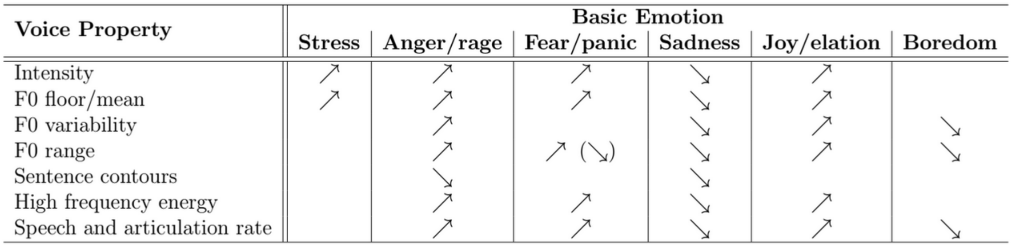
2.2 Robots and Embodiment

It was found through prior research that people empathize (Seo et al. 2015) and are more engaged with an embodied robot.(Lee et al. 2006) (Kidd & Breazeal 2005) (Kiesler et al. 2008) Physical embodiment can make a difference in perception of a social agent's capabilities and the user's enjoyment of a task.(Wainer et al. 2006) Nourbakhsh et al. created a robotic tour guide that expressed emotion vocally.(Nourbakhsh et al. 1999) Participants felt a greater sense of presence, felt it was more lifelike, and disclosed less private information with an embodied robot versus an avatar.(Kiesler et al. 2008) Based on a survey on experiments with embodied robots, telepresent robots and avatars, people preferred an embodied robot. People felt a higher level of arousal, responded more favorably, had a stronger response, and found physically present robots more persuasive. (Li 2015)

In the study for this project, the emotive response of subjects to a PC were compared to responses to a mobile robot and to a static robot. It was hypothesized that a robot would create a stronger emotive response because the robot would be more engaging and exciting.

2.3 Theories of Emotion

There are three major ways to look at emotion. (Davidson et al. 2003) Emotions can be viewed as discrete, dimensional, or appraisal-based. Ekman popularized theories about discrete emotion. (Ekman & Friesen 1969) The discrete categories model has no biological backing, and does not allow for a range of emotions or mixed emotions. Appraisal-based emotion theory is the theory that emotions are derived from people's expectations and interpretations of an interaction without the need for arousal. Appraisal based emotion theory is incredibly hard to categorize because it is descriptive rather than categorical,and it is rarely used when looking at emotion and sound. While discrete categories of emotions are frequently used for studies, A dimensional model was chosen instead for the study because emotional responses to sound are frequently more subtle and complex than the expression of a single emotion. A theory of emotion using a dimensional model can cover and express a range of emotions and mixed emotions. Dimen-



Voice Property	Basic Emotion					
	Stress	Anger/rage	Fear/panic	Sadness	Joy/elation	Boredom
Intensity	↗	↗	↗	↘	↗	
F0 floor/mean	↗	↗	↗	↘	↗	
F0 variability		↗		↘	↗	↘
F0 range		↗	↗ (↘)	↘	↗	↘
Sentence contours		↘		↘		
High frequency energy		↗	↗	↘	↗	
Speech and articulation rate		↗	↗	↘	↗	↘

Figure 2.5: An increased valence is correlated with a higher pitch, higher deviation, larger range, higher mean intensity, larger intensity deviation, faster speech rate, shorter syllable duration and shorter/less frequent pausing. Decreased valence is correlated with the opposite effects. (Scherer et al. 2003)

sional models have been linked to levels of chemicals and neurotransmitters in the brain, with distinct biological pathways.(Lövheim 2012) Most models use pleasure and arousal as the two major axes.(Russell 1980) Dominance/submission is a well accepted and commonly used 3rd dimension creating the PAD (Pleasure, Arousal, Dominance) space. (Russell & Mehrabian 1977) The pilot study for this project used the PAD model for studying user's response.

2.4 Emotion and Prosody

This project focused on designing a generative sound system for creating emotive sound. The sounds generated by Read's system was not recognized as consistent emotion and it ignored major/minor mode and steady state versus percussive sounds. (Read 2014) Major/minor mode has been found to relate to valence.

(Turner & Huron 2008) Huron found that instruments that were more percussive were believed to be unable to express sadness. (Huron et al. 2014) In Le Groux's pilot study on timbre, valence was not correlated with any emotion, but the study only used percussive sounds. (Le Groux & Verschure 2010) Valence has also been correlated with both pitch, intensity and rate in speech. While an increased valence is correlated with a higher pitch, higher deviation, larger range, higher mean intensity, larger intensity deviation, faster speech rate, shorter syllable duration and shorter/less frequent pausing. Decreased valence is correlated with the opposite effects. (Scherer et al. 2003)

The studies that were reviewed fell into 2 categories, those using computers, particularly spectral analysis, and those using people to analyze their data. Through the literature review, the audio aspects that were studied mainly fall into 4 categories: F0/Pitch, Amplitude/Intensity, Speech Rate/Tempo, and Articulation/Timbre. Because research from both the music and linguistics background was studied, some terms were more scientific and measurable while other terms were more vague and could be interpreted in different ways. This also meant that different vocabulary was used in different studies to describe the same idea. (see Table 2.1- 2.4) F0 (5), F0 mean (5), F0 perturbation/range (5), F0 variability (3), F0 contour (3), high frequency-energy (2), pitch (8), pitch average (1), pitch range (3), pitch variation (3), pitch maximum (1), and major/minor mode (5)

Table 2.1: F0/Pitch in Research

Type	Citation
F0	(Banse & Scherer 1996) (Tartter 1980) (Ohala 1996) (Ohala 1980) (Ohala et al. 1997)
F0 mean	(Williams & Stevens 1972) (Sobin & Alpert 1999) (Johnstone et al. 2001) (Banse & Scherer 1996) (Scherer et al. 2003)
F0 perturbation/range	(Williams & Stevens 1972) (Johnstone et al. 2001) (Banse & Scherer 1996) (Scherer et al. 2003) (Cowie et al. 2001)
F0 variability	(Sobin & Alpert 1999) (Johnstone et al. 2001) (Banse & Scherer 1996)
F0 contour	(Johnstone et al. 2001) (Banse & Scherer 1996) (Cowie et al. 2001)
high frequency-energy	(Banse & Scherer 1996) (Johnstone et al. 2001)
pitch	(Streeter et al. 1983) (Ohala 1983) (Scherer et al. 1973) (Lieberman & Michaels 1962) (Huron et al. 2014) (Apple et al. 1979) (Cowie et al. 2001) (Huron et al. 2006)
Pitch average	(Murray & Arnott 1993)
Pitch range	(Murray & Arnott 1993) (Huron 2008) (Cowie et al. 2001)
Pitch Variation	(Scherer et al. 1973) (Murray & Arnott 1993) (Breitenstein et al. 2001)
Pitch Maximum	(Schutz et al. 2008)
Major/ Minor Mode	(Turner & Huron 2008) (Schutz et al. 2008) (Dalla Bella et al. 2001) (Post & Huron 2009) (Huron 2008)

Table 2.2: Amplitude/Intensity in Research

Type	Citation
Amplitude	(Tartter 1980) (Sobin & Alpert 1999) (Lieberman & Michaels 1962) (Streeter et al. 1983)
Energy	(Scherer et al. 2003) (Johnstone et al. 2001) (Banse & Scherer 1996)
Loudness	(Scherer et al. 1973) (Siegman & Boyle 1993) (Huron et al. 2014)
Intensity Mean	(Banse & Scherer 1996) (Cowie et al. 2001) (Murray & Arnott 1993)

Table 2.3: Timbre/Articulation in Research

Type	Citation
Articulation	(Murray & Arnott 1993)
Timbre	(Schutz et al. 2008) (Hailstone et al. 2009) (Huron et al. 2014) (Huron et al. 2006)
Voice Quality	(Murray & Arnott 1993) (Cowie et al. 2001)

Table 2.4: Speech Rate/Tempo in Research

Type	Citation
Speech Rate/Tempo	(Siegman & Boyle 1993) (Scherer et al. 1973) (Johnstone et al. 2001) (Murray & Arnott 1993) (Banse & Scherer 1996) (Apple et al. 1979) (Breitenstein et al. 2001) (Huron et al. 2014) (Dalla Bella et al. 2001)
Duration	(Williams & Stevens 1972) (Tartter 1980) (Sobin & Alpert 1999) (Schutz et al. 2008) (Scherer et al. 2003) (Cowie et al. 2001)
Pausing Total Time	(Sobin & Alpert 1999) (Scherer et al. 1973) (Cowie et al. 2001)

were commonly studied in papers. F0 median, range, whether the key is major or minor, and contour were chosen as variables for this project's pilot study. Amplitude (2), energy (3), loudness (3), and intensity mean (3) were the most common ways to look at amplitude. Sustain mean, sustain variance, attack and sustain difference, and contour were used to describe the amplitude. A portion of this is also used for the envelope of each note, which expresses the articulation or timbre of the sound. Articulation (1), timbre (4) and vocal quality (2) were used to describe sound. The envelope is described using an ADSR envelope, which has an attack, decay, sustain mean, sustain variance and release. Speech rate/tempo (9), duration (6), and pausing total time (3) were analyzed to describe speech

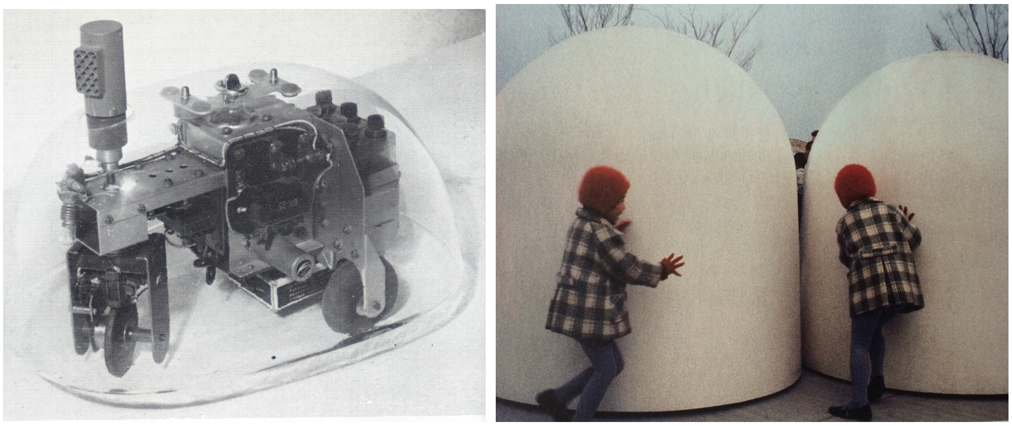


Figure 2.6: Grey Walter's tortoises, named because of their shape and slow speed. (Sutherland 1960) Robert Breer, "Floats" (left) at the Pepsi-Cola-Pavilion, Osaka 1970 Roy Lichtenstein Foundation, Photo: Shunk Kender

rate/tempo. Pausing frequency median and range, as well as length mean and variance were also used.

2.5 Robots in Media Arts

2.5.1 Cybernetics

ROVER references many early works of cybernetic art. By using an iRobot CREATE, ROVER references Grey Walter's tortoises and Robert Breer's sculptures at Pepsi Pavilion. Some of the first autonomous robots were Elmer and Elsie, by Grey Walter, constructed in 1949. (see Figure 2.6) They were called tortoises because of their shape and slow speed. Their functionality is similar

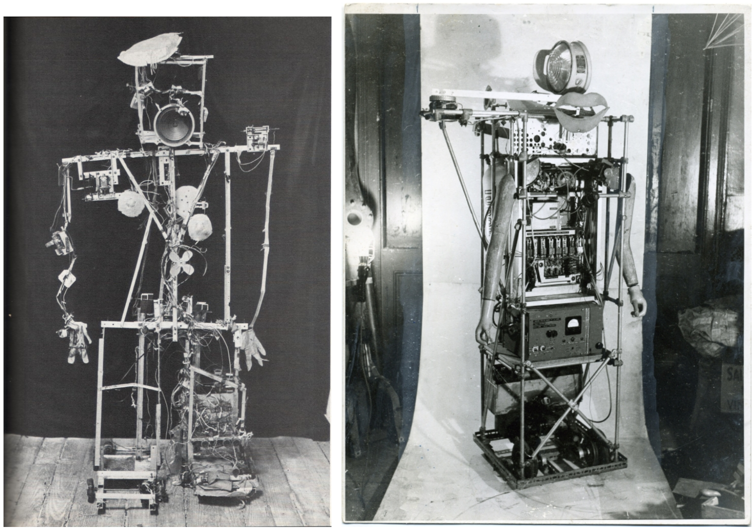


Figure 2.7: Robot K-456, Nam June Paik 1965 (Photo: Peter Moore) and ROSA BOSOM, Bruce Lacey 1964 (Photo: Bruce Lacey) (Hoggett 2010)

to today's Roomba, with bump sensors to avoid obstacles, and Elsie was able to return to a docking station to recharge when running low on power. (Walter 1950) Robert Breer's sculptures at the Pepsi Pavilion in 1970 were 6 feet high and emitted sound while moving around at less than 2 feet per minute. (see Figure 2.6) (Prade 2002) ROVER is particularly reminiscent of these sculptures, a 6 foot tall robot moving slowly and emitting sound.

These early autonomous robots did not interact with the public. To create an engaging robot, many early robots were remote controlled and were used for disruption like Bruce Lacey's ROSA BOSOM (Radio Operated Simulated Actress Battery Or Standby Operated Mains) with Mate (Reichardt 1969) and K-456 by

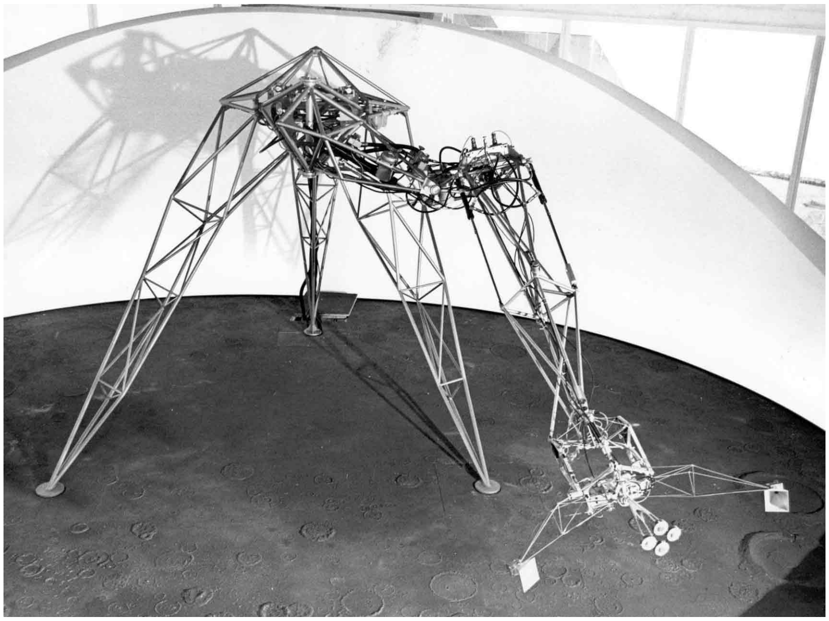


Figure 2.8: The Senster, Edward Ihnatowicz 1970 (Photo Credit: The Philips Archive 1971) (Zivanovic 2007)

Nam June Paik and Shuya Abe. (see Figure 2.7) K-456, built in 1964, was a 20 channel radio controlled robot originally considered "androgyne" but cast as female in the United States. "Robot-K456 can bow, walk, give a speech (recorded by the then Mayor-elect of New York, John Lindsay), lift each arm independently and wiggle its representational torso. It also defecates on the floor of the gallery by remote control. Paik's robot looks mechanically unreliable and he admits that it needs constant attention." (*Electronic Design* 1966) Rosa Bosom was originally designed as an actress to play the Queen of France in the production of Three Musketeers, at the Arts Theatre 1966. (Reichardt 1969) These robots were shown at Cybernetic Serendipity with Gordon Pask's Colloquy of Mobiles, robotic mobiles which had very simple interaction tasks to interact with each other using light. The viewers were given flashlights so that they could interact with the robots. (Reichardt 1969)

In 1970, The Senster, by Edward Ihnatowicz, was the first computer-controlled robotic sculpture that was interactive with the public, using 4 microphones and 2 doppler radar arrays. (see Figure 2.8) It was an 8ft tall and 15ft long hydraulically activated sculpture that followed the sound and motion of the spectators. It was attracted to sound and movement but avoided loud noises and quick movement. There was a 5 degree of freedom arm which was novelty at the time. (Benthall 1972)

2.5.2 Contemporary Robotic Art

Robots Emulating Humans

Berenson, named after Bernard Berenson, is a robot art critic by anthropologist Denis Vidal and robotics engineer Philippe Gaussier. (see Figure 2.9) The critic observes viewers' reactions to art and learns what is “good” and “bad” art. Then he moves toward art works that are “good” and smiles at them, and frowns at “bad” art. This robot uses a neural network to learn. (Pangburn 2016) While there are few robots that compose music, drawing robots and machines

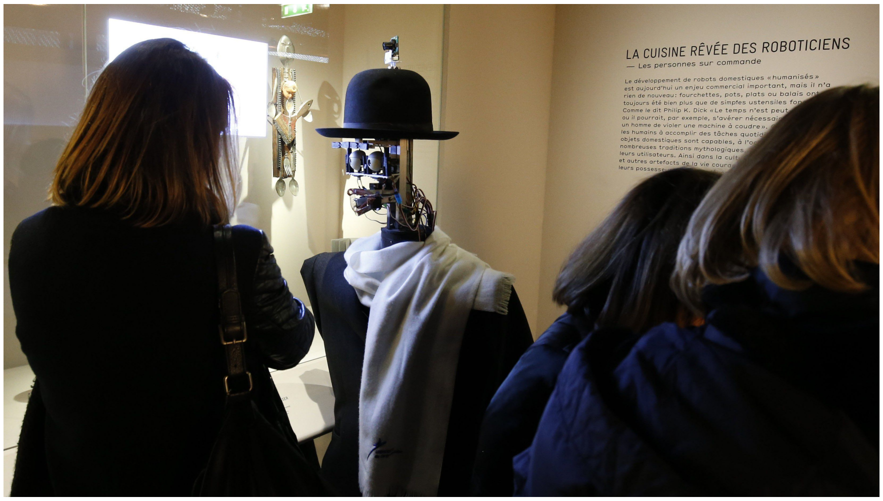

LA CUISINE RÊVÉE DES ROBOTICIENS
— Les personnes sur commande
Le développement de robots domestiques «humanisés»
est aujourd'hui un enjeu commercial important, mais il n'a
rien de nouveau: fourchettes, pots, plats ou balais ont
toujours été bien plus que de simples ustensiles fon
Comme le dit Philip K. Dick «Le temps n'est peut
ou il pourrait, par exemple, s'avérer nécessai
un homme de violer une machine à coudre».
les humains à accomplir des tâches quotidi
objets domestiques sont capables, à l'o
nombreuses traditions mythologiques.
leurs utilisateurs. Ainsi dans la cult
et autres artefacts de la vie coura
leurs possesse

Figure 2.9: Berenson, named after Bernard Berenson, is a robot art critic by anthropologist Denis Vidal and robotics engineer Philippe Gaussier. (Pangburn 2016)

are pervasive in art and have been explored by Shih Yun Yeo, Bálint Bolygó, Patrick Tresset, Nils Völker, Jen Hui Liao, Brian De Rosia, Guy Ben-Ary, Harold Cohen, Jörg Lehni, Jeff Badger, and Fernando Orellana. (Yeo 2015) (Bolygó 2015) (Tresset & Leymarie 2013) (Völker 2009) (Debatty 2009) (DeRosia 2014) (Bakkum et al. 2007) (Cohen 1995) (Lehni 2008) (Badger 2008) (Orellana 1999) Louis-Phillipe Demer's Tiller girls is a live interpretative performance with simple robots that can only move their necks and waists. (Demers 2015)

Robotic instruments have existed since the 14th century with the Carillon. (Leichtentritt 1934) Mechanical music was part of the Dadaist movement, an example is the Ballet Mécanique Dadaist film by George Antheil. (Léger & Murphy

1924) The music was performed by LEMUR (League of Electronic Musical Urban Robots) at the National Gallery in 2005 with a computer driven robotic ensemble. (Lehrman & Singer 2008) Peter Ablinger made a piano speak in "Speaking Piano." (Ablinger 2006) In the pilot study for this project, ROVER's final frequency range was based on a piano.

The relationship between sound and the human body is explored in CodAct's Pendulum Choir, where the performers work with a system to create sound. (Cod.Act 2011) Stelarc explored the relationship between man and machine in many of his works, both augmenting and extending his body. (Atzori & Woolford 2015) Robots have been built that reproduce bodily functions like Kevin Grennan's robot that sweats (Grennan 2011) and Alexitimia, Paula Gaetano's sweating robot. (Adi 2008) Another example of this is Cloaca, Wim Delvoye's machine that defecates, and the previously mentioned K-456 by Nam June Paik. (Criqui 2001) (*Electronic Design* 1966)

The first speaking machine, a person controlled mechanism with bellows, was designed in 1769 by Wolfgang von Kempelen. It could only say a few words. (Kempelen et al. 1970) Around the same time, C. G. Kratzenstein constructed various shaped tubes that produced five vowel sounds. The first electrical speaking machine was the Voder designed by Homer Dudley in 1939. (Dudley et al. 1939) Research has continued to the present with work like the Waseda Talker Series from the Humanoid Robotics Institute at Waseda University (Fukui et al. 2006), and Hideyuki Sawada's KTR-2 which sings. (Sawada 2007) The first computer to sing was the IBM 7094 in 1961, singing the song Daisy Bell. Vocals programmed by John Kelly and Carol Lockbaum inspired a scene in 2001: A Space Odyssey. (Smith 2010)

Autonomous Robots

Autonomy was an important drive when creating ROVER. Jed Berk's ALAVs (Autonomous Light Air Vessels) are a flock of floating, sheeplike balloons that can communicate with lights and movement. In the first version, viewers could befriend the ALAVs and change the flocking behavior by feeding them. In later versions, viewers could communicate with the ALAVs via cell phone and their choices of being friend or foe affected the ALAV flock's actions. (Berk 2009) Robots have moved outside of the gallery with Theo Jansen's Strandbeests, which he is trying to make completely self sufficient on the beach. (Jansen 2008) Fernando Orellana's work, Elevator's Music, 2007, has robots driven by sound that hide in the elevator. (Orellana 2007) Gilberto Esparza's Urban Parasites' "... intention is to create life forms that exist at the expense of energy sources generated by the human species, which can be found in the urban environment." (Esparza 2007) Art also extends to remote environments, exploring places where there is little

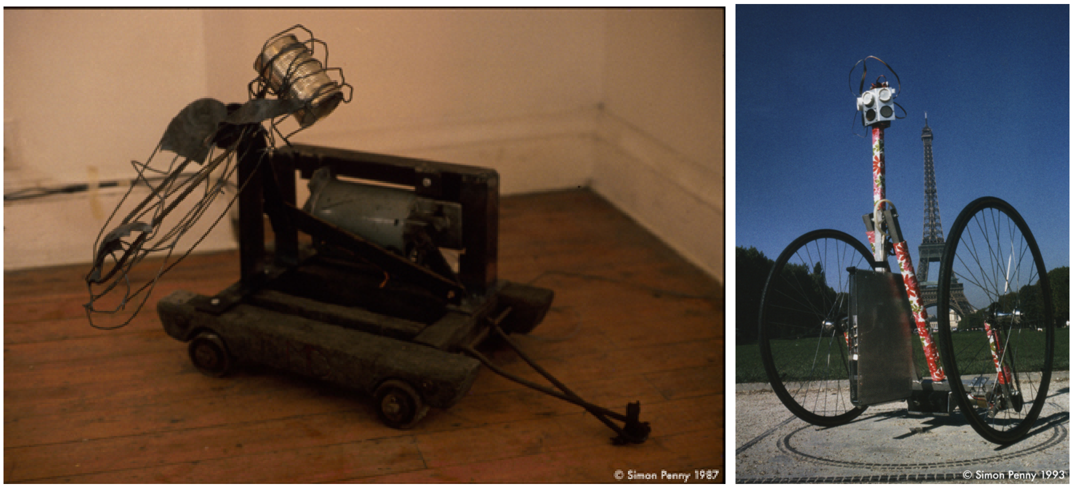


Figure 2.10: Simon Penny's Stupid Robot, 1985 (left) and Petit Mal, 1993 (right) (Penny 2011)

to no human intervention. Michael Snow's film La Région Centrale was created completely with a mechanical camera surveying a remote area of Canada. (Snow 1972) A strong counter-cultural streak exists in making robots such as the work done by the Survival Research Laboratory. (Pauline 1979)

Case Study: Simon Penny

Simon Penny explores human interaction with technology and robotic recreation of human behaviour. His first work on these topics was Stupid Robot in 1985. (see Figure 2.10) Stupid Robot was designed to be reminiscent of a legless beggar, and it shook a can of metal scraps when approached. In 1990 he created a heat-seeking anti-personnel sculpture called Pride of Our Young Nation. Pride of Our Young Nation was designed to look like an artillery cannon and to use an infrared heat sensor to aim at its victims. Once it found its victims, it would “fire” by rotating a large metal cone covered in spikes towards them. Petit Mal, is an autonomous interactive robot, designed to be more simple than functional. (see Figure 2.10) Its basis is a pendulum and two bicycle wheels. It uses ultrasonic and piezoelectric sensors to navigate the space and find people, which it then follows. It is adorable in its clunkiness. In Phatus, Penny works with the idea of trying to reproduce how people make noises by creating artificial vocal cords and lungs. He is still currently working on this project, another form of interaction, speech, but instead of just using electronic forms of synthesis, this speech is purely mechanical. (Penny 2011)

Case Study: Ken Rinaldo

Ken Rinaldo explored sound as communication for robots and autonomous robots that photographed and interacted with the public in a way similar to ROVER. Ken Rinaldo’s early sculpture, Cyber-squeaks, 1987, was a series of small hanging sculptures that reacted to touch and light by emitting sound. (see Figure 2.11) Ken Rinaldo continued to explore sound and interaction with the Flock in 1994. Flock was three robotic hanging arms that interacted with the public through movement and communicated with each other using telephone tones.

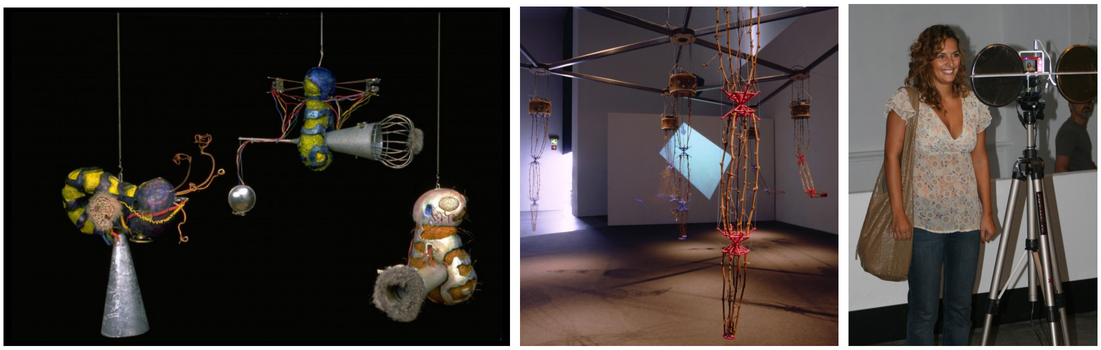


Figure 2.11: Ken Rinaldo's work left to right: Cyber-squeaks 1987, Autopoiesis 2000, Paparazzi Bots 2009 (Rinaldo 2015)

They sensed the environment with microphones and infrared sensors. This was later expanded into Autopoiesis, the actions of which evolved based on interactions with the public. (see Figure 2.11) Rinaldo first started exploring autonomous robots with Augmented Fish Reality, where Beta fish could control mobile tanks which moved around the space. This is similar to Garnet Hertz's Cockroach Controlled Robot in 2008. (Hertz 2008) Rinaldo's next autonomous robot was the Paparazzi Bots, a series of human-height robots that would move toward people and take pictures of them like paparazzi. (see Figure 2.11) They moved at human speed, avoiding obstacles using multiple microprocessors, cameras, sensors, and a custom rolling platform. (Rinaldo 2015)

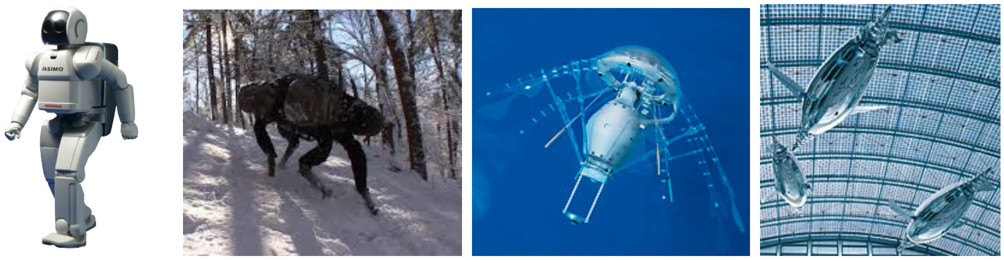


Figure 2.12: Robots from industry, left to right: Asimo, BigDog, AquaJelly, AirPenguin (Sakagami et al. 2002) (Raibert et al. 2008) (Festo & Co 2009) (Fischer 2009*a*)

2.6 Robots in Research/Industry

Biologically inspired robots are found throughout research in locomotion. (see Figure 2.12) Outside of more standard humanoid robots like Asimo,(Sakagami et al. 2002) examples of animal inspired robots are Boston Dynamics' research with the Wildcat, Big Dog (Raibert et al. 2008) and Cheetah (Sapaty 2015) and Festo's aqua penguins and aqua jelly. (Fischer 2009*b*) (Festo & Co 2009) Complex locomotion is rarely used in home robotics, where Roomba style locomotion is more reliable.

Pepper by SoftBank Mobile and Aldebaran Robotics SAS is a Japanese home robot in which can read emotion. (see Figure 2.13) Researchers are already investigating its use for teaching via telecommunication. (Tanaka et al. 2015) Pepper costs approximately 1800 dollars and is currently only available in Japan. While the goal of Pepper is to make people smile, there has been little research

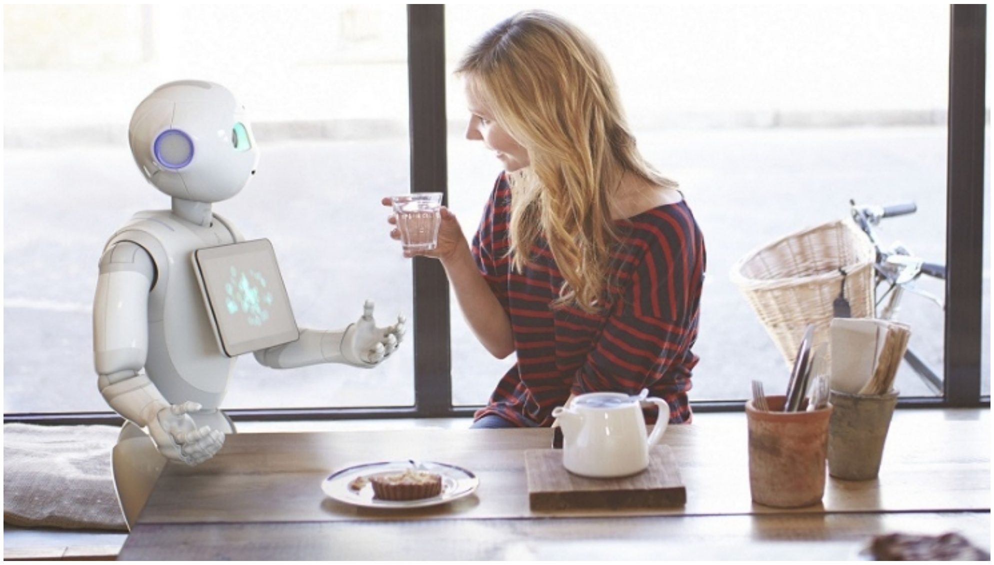


Figure 2.13: Pepper by SoftBank Mobile and Aldebaran Robotics SAS is a Japanese home robot which can read emotion.(Aldebaran 2015)

done on how emotive Pepper is. (Greer 2014) JIBO is a personal robot yet to be released by Dr. Cynthia Breazeal's new start-up. (see Figure 2.14) JIBO is a stationary tabletop robot advertised as a personal assistant like SIRI or Cortana that will be able to track emotions. Again there is little research yet on the emotional range of JIBO. (Rane et al. 2014) Both Pepper and JIBO can be used for telepresence.

While industrial robots have been around since 1961, (Bekey & Yuh 2008) personal robots first appeared in the early 1980s. Their main function was educational, to teach people how robots work. (Bell 1985) The first service robot that had a job was the HelpMate service robot, a robotic courier for hospitals, in 1988.

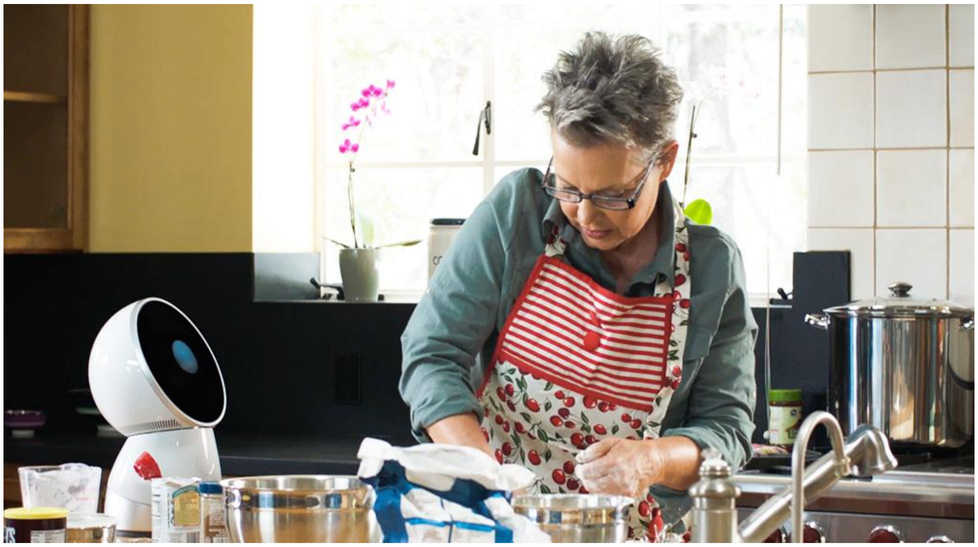


Figure 2.14: JIBO is a personal robot yet to be released by Dr. Cynthia Breazeal's new start-up. (Jibo 2015)

(Evans et al. 1989) Personal helper robots are nowhere near the point where they can interact in non-controlled environments. (Royakkers & van Est 2015) Care robots for the home, along with social service robots in museums and restaurants have begun to appear in the last decade. (Pieskä et al. 2012)

"Socially assistive robots describes a class of robots that is the intersection of assistive robotics (robots that provide assistance to a user) and socially interactive robotics (robots that communicate with a user through social and non-physical interaction)."(Feil-Seifer & Matarić 2011) Paro, the robotic seal, is one example of a socially assistive robot. Pets are seen as therapeutic for the elderly and sick, but it is hard for these groups to maintain the responsibility of having a pet. Paro interacts with simple sounds and movements, responding to being petted and held. There have been no studies specifically on the sounds that Paro makes. (Feil-Seifer & Matarić 2011) Mamoru is another socially assistive robot that takes note when people take their medication and makes sure that they don't take it twice. (Wu et al. 2012) Other examples of socially assistive robots are the Nursebot project (Pollack et al. 2002), Robocare project (Bahadori et al. 2003) and Care-o-bot(Graf et al. 2004).

HANC was an early healthcare robot in 1995 that reminded patients to take their pills and could run routine tests. (Kaufman & Van Ellin 1995) InTouch Health was one of the first in the telepresent robot market using them in rehabilitation centers, eldercare facilities and hospitals.(Wang et al. 2005) Roomba produced the CoWorker robot in 2002 and ConnectR robot in 2007, though neither of them were a commercial success. (Tsui et al. 2011) Telepresent robotics is a field that is undergoing rapid expansion in research, office, eldercare and healthcare with PRoP, Giraff, QB, Texai, Beam, VGo, PEBBLES, MantaroBot, Double, mObi, Jazz Connect, iRobot Ava, 9th Sense Helo and Telo, RP-7 and MeBot. (Kristoffersson et al. 2013) Telepresent robots have expanded into remote environments like undersea and space exploration. (Hine et al. 1994)(Corbett et al. 2012)

Chapter 3

Implementation

The construction of ROVER was an iterative process. The project was exhibited in the Media Arts and Technology End of Year Show in 2013, 2014 and 2015 and a pilot study was run using it. Table 3.1 is an overview of the evolution of the project.

3.1 Precursors to ROVER

3.1.1 Tracking/Spatial Audio

The concept of ROVER was initially conceived in the transLAB in the Fall 2012. Retro-reflective sphere markers were attached to a remote control All Terrain Vehicle so it could be monitored by the tracking system. A computer program was written to keep the vehicle inside of the tracked area by taking into account where it was and using an Arduino to press the buttons on the remote control.

Table 3.1: Implementation overview

Version	Sound	Sensors	Aesthetics	Brains
ROVER	Genetic Algorithm producing series of Midi notes played by the iRobot Create	-iRobot Create sensors -Melexis 90620 -GoPro Camera	Furry 5 foot tall robot	Arduino Mega connected to computer via xBee
ROVER 2.0	Genetic Algorithm producing series of Midi notes played by the iRobot Create	-iRobot Create sensors -Melexis 90620 **-Raspberry Pi Camera Module** **-Sharp Proximity Sensors**	**6 foot tall acrylic structure**	Arduino Mega **Connected to a Raspberry Pi**
ROVER 3.0	Genetic Algorithm for **Parameters based on Prosody research using speakers**	-iRobot Create sensors -Melexis 90620 -Raspberry Pi Camera Module -Sharp Proximity Sensors	6 foot tall acrylic structure	Arduino Mega Connected to a Raspberry Pi
Pilot Study	**Two sets of parameters that define “happy” and “sad” sounds based off of Prosody research** using speakers	-iRobot Create sensors -Melexis 90620 -Raspberry Pi Camera Module -Sharp Proximity Sensors	6 foot tall acrylic structure	Arduino Mega Connected to a Raspberry Pi **talking to a server hosting a web interface for data collection**

There was also an LED array that lit up representing the location of the vehicle. The sound was produced through an interactive real time program that took field recordings and spatialized them based on ROVER's location, using an 18.1 speaker array. This installation was reminiscent of the Mars Rover in a black room on a grey tarp. This project focused on sound and its relationship to an object in space. After finishing the project, the next step was to take it outside of the transLAB and make it autonomous.

3.1.2 Facial Tracking and Analysis

Winter of 2013 a project using facial tracking was created, in Python using OpenCV. The program used haar-cascades to find faces and eigenface analysis to determine whether or not the person had been seen before. Color and blob detection were used to track people even if their faces weren't visible. It was decided only to use face detection, and not the other more complex processes in ROVER because they were very processing-intensive.

3.2 ROVER (Concept)

ROVER was originally conceived in Marko Peljhan's WORLDCHANGING class. In the prototype reactivity, spatial location, and sensing were paramount.

The spatial terrain of the robot was explored and the locational activity was sensed and displayed through data visualization. Both the Systemics and Experimental Visualization (ExpVis) Lab were involved in the process. In the Systemics Lab the focus was on mobility and reactive systems, while in the ExpVis lab the mapping and heat sensing data were visualized. ROVER was used to map the space, where people were and where ROVER interacted with people. The map could be used by ROVER to find people in a space. The intention was to give ROVER the aspects of intelligence, knowing its space and entities in the space. ROVER sensed the space through bump sensors, sensed people in the space through heat and video analysis, and reacted to the space by avoiding obstacles, moving toward warmth and playing a series of notes when a person was seen.

The goal was to create a "puppy" for the hallway of the second floor of Elings Hall, a robot that would find people, track them, recognize them and greet them. To address the cold desolate feeling of the hallway, ROVER used heat detection to find the warmest body in the room. The idea of using smile detection as positive reinforcement would create a friendly interaction. Originally the plan was to have ROVER react through movement, however it was later realized that if ROVER moved, it might not detect the viewer and therefore could not receive feedback. It was then decided to use audio as a form of communication instead, and the performance was modified to see if ROVER's learning algorithm would converge on the most smile producing song.

3.2.1 Interaction Design

ROVER wandered the exhibition, moving toward the nearest detectable heat source. When a face was detected using haar-cascades, ROVER stopped and played a simple song created by a genetic algorithm. ROVER detected how much the person smiled during the song. The amount the person smiled was fed into the genetic algorithm as the fitness function output. The song was made up of 5 midi notes played by the Roomba. As an end result, ROVER would create the best song that made people happy. (see Figure 3.1)

Sound Generation Algorithm

Songs were generated using midi notes sent to the Roomba. Each song contained of a series of 5 notes. which were generated through a genetic algorithm. There were 7 midi notes, 45 (A, 110Hz), 47 (B, 123.5Hz), 48 (C, 130.8Hz), 50 (D 146.8Hz), 52 (E, 164.8Hz), 53 (F, 174.6Hz), and 55 (G, 196Hz). The initial pool was all 7^5 options. A test population of 50 was selected from the initial pool, randomly using Python's random.sample function. Each song in the test population was played to a viewer and the amount the viewer smiled in response

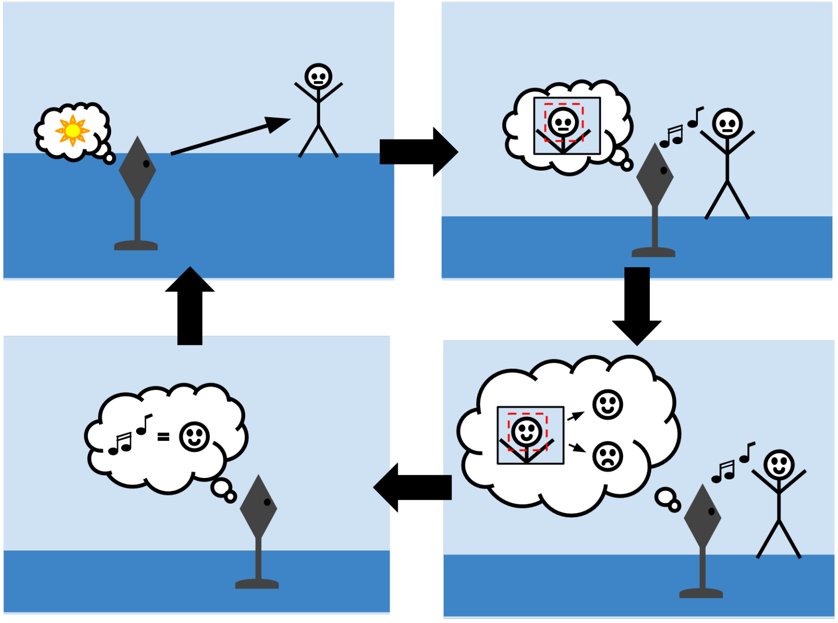


Figure 3.1: ROVER searches the space for heat (upper left), stops and sings when it detects a face (upper right), measures how much the person is smiling during the song (lower right), uses the song and reaction to learn. (lower left)

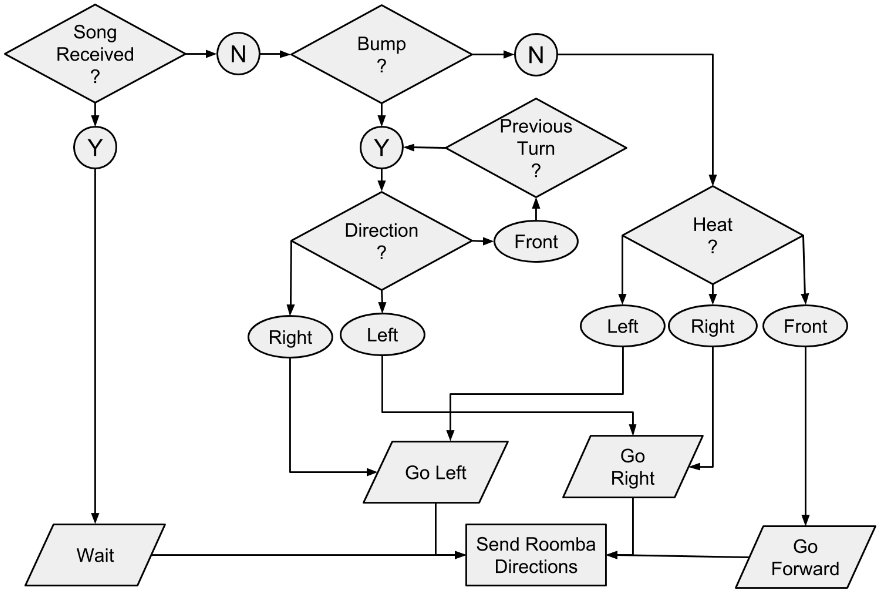


Figure 3.2: ROVER movement diagram

would be recorded. The amount the viewer smiled was calculated using a haar cascade trained on smiles from OpenCV. The top 20 songs that people smiled to the most were considered as parents for the next generation. The parents then created 30 offspring via crossover, in which 15 pairs of parents swapped a portion of their song to create 2 different songs. Then each child had a 10% mutation probability in which they were mutated by a random gaussian distribution from 0 with a standard deviation of .5. This final population of parents and children then became the next test population.

Movement Algorithm

The movement algorithm first checked to see if a song was received. (see Figure 3.2) If it was, the Roomba would stop and sing a song. If there was no song, the Roomba would check to see if it had run into anything. If it had run into something, it would move to avoid that object; otherwise it would move toward the warmest direction.

3.2.2 Technology Stack

The technology used was the iRobot Create, Arduino Mega, xBee, Melexis 90620, GoPro with a wifi backpack and a laptop. (see Figure 3.3) The original intent was to use a Raspberry Pi but the camera module wasn't released until May

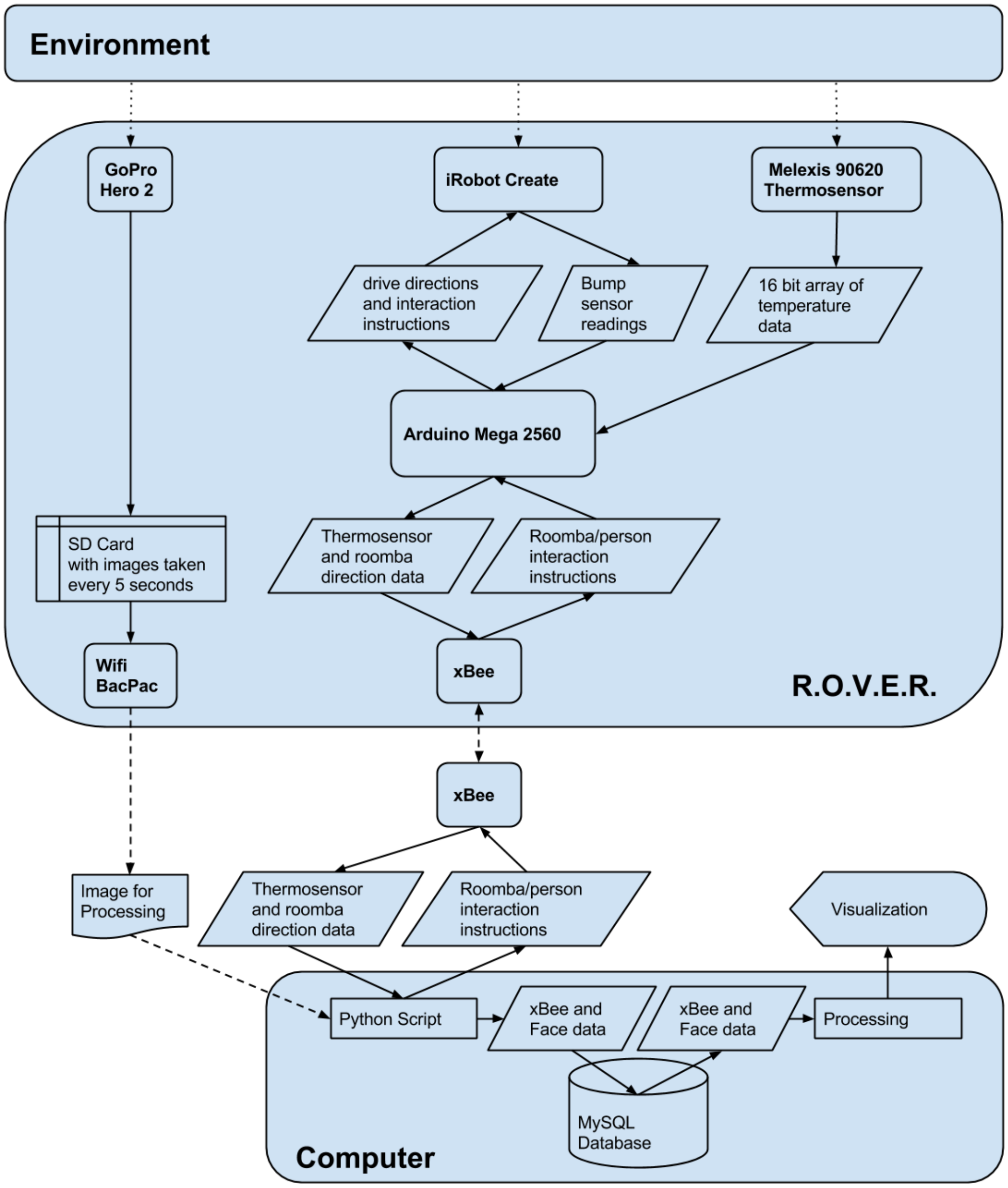


Figure 3.3: ROVER technology stack

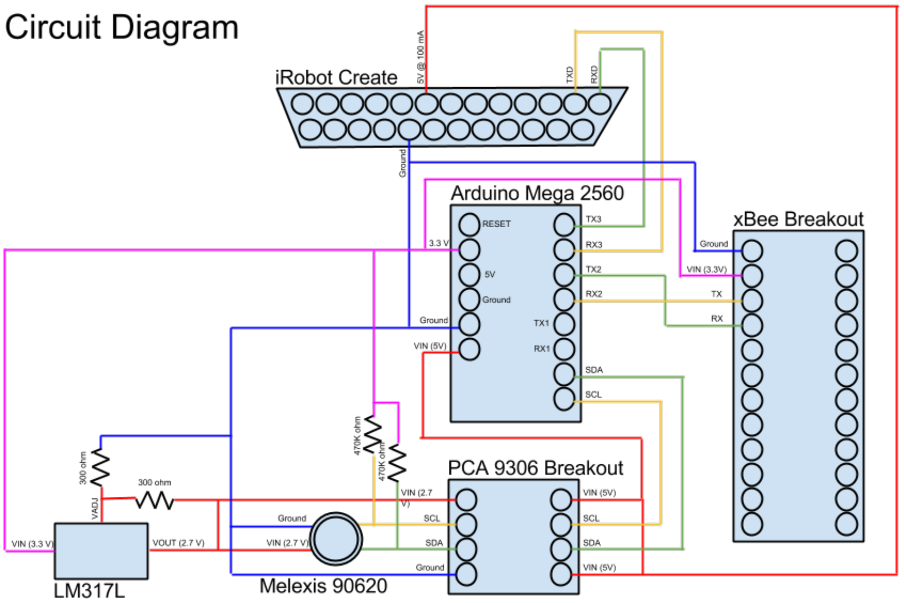


Figure 3.4: Wiring diagram for ROVER

2013. To compensate, a GoPro with a wifi backpack was used to take pictures and a laptop connected to the wifi network pulled the pictures that were taken. (see Figure 3.4 for the wiring diagram)

3.2.3 Structure and Aesthetic Design

ROVER's visual appearance was heavily influenced by the skeletal form. (see Figure 3.5) ROVER had a large foot or base, a very thin long neck or leg, and a large diamond shaped head or body which housed the electronics. To be able to see the viewers, the cameras and sensors needed to be at face level. To cover the

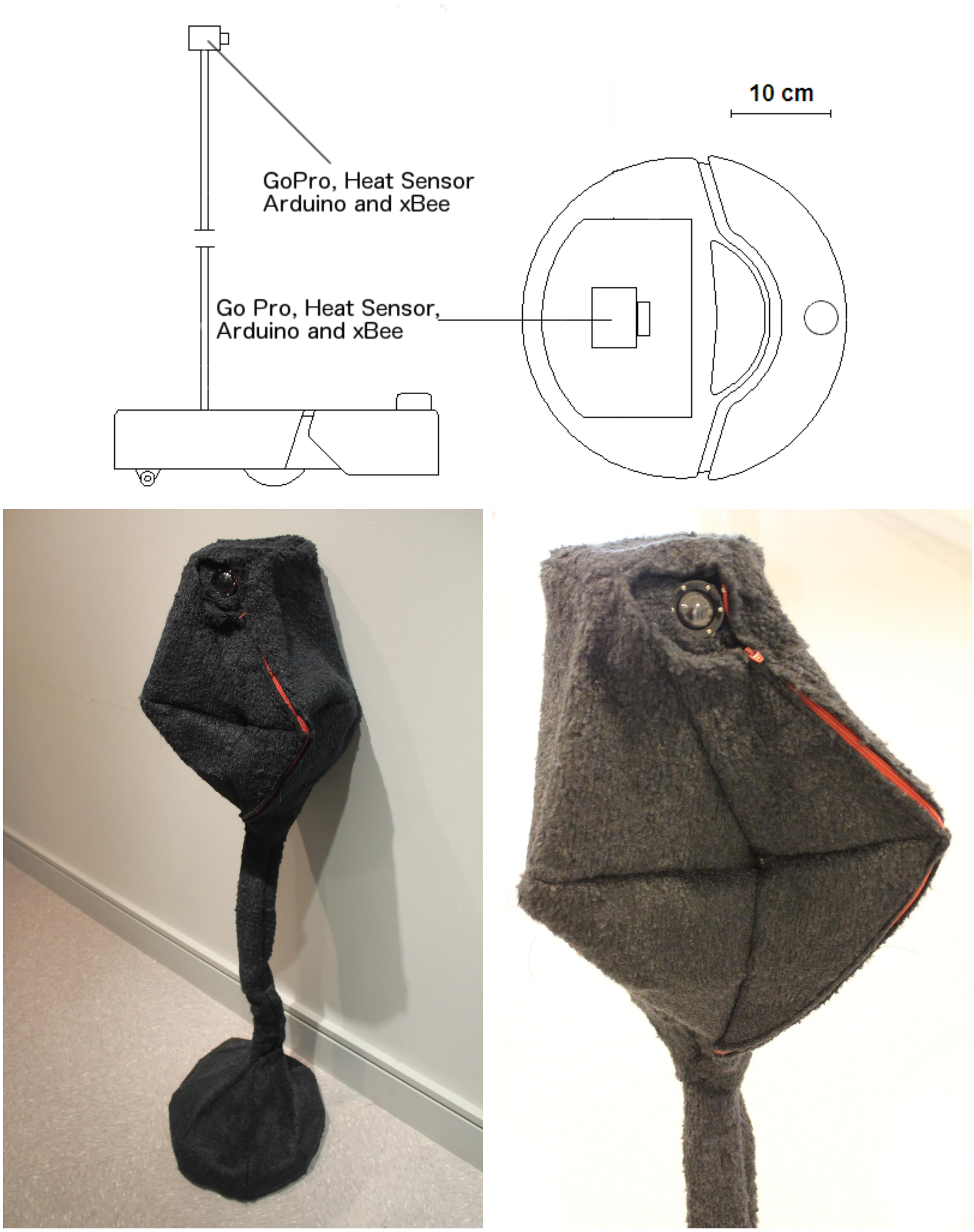


Figure 3.5: Basic structural design and final appearance of ROVER

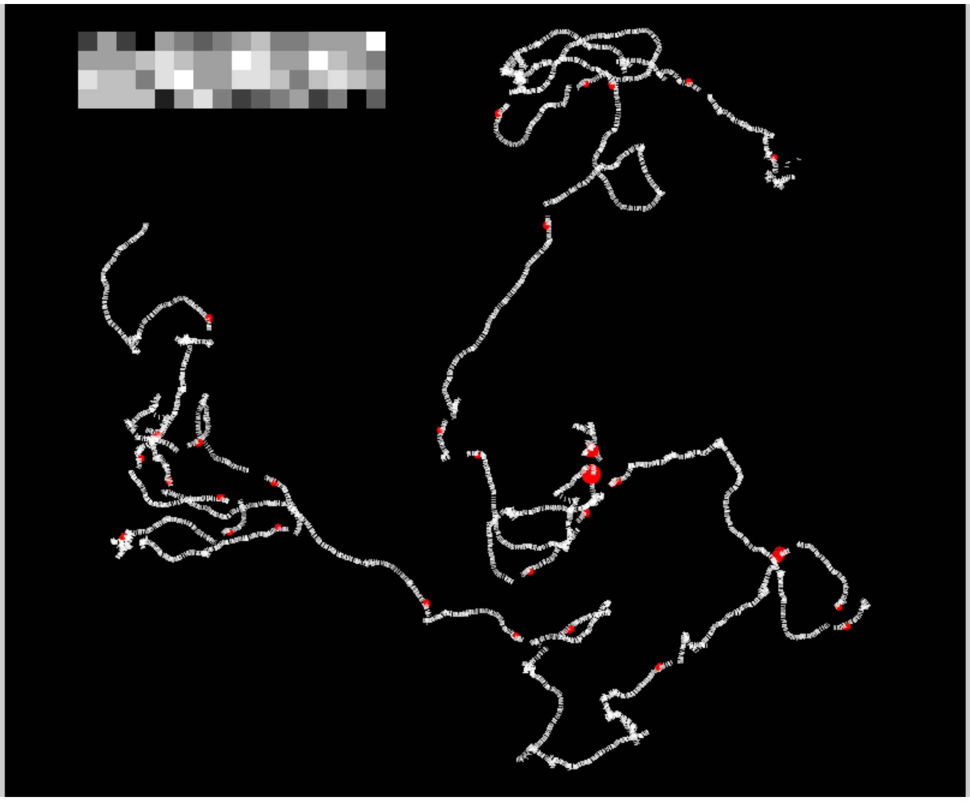


Figure 3.6: Realtime visualization of data from ROVER

skeletal form, a dark grey-blue felt cover reminiscent of fur was created to give ROVER a puppet-like quality. (see Figure 3.5)

3.2.4 Processing Visualization

The visualization of ROVER's interaction and movement took the last 3000 points and mapped them. (see Figure 3.6) For each point the 64 bit temperature sensor array was displayed. The left and right wheel velocities were converted to a trajectory, based on time between each data point. If a face was seen at that point, a circle was drawn next to the heat grid. The circle's radius was based on how many faces were visible. In the upper left corner was a larger, real time grid of the last heat sensor data received, with a red circle next to it if a face was seen. The visualization was updated for each frame with the newest data. The visualization was written in Processing, pulling from the Roomba data. At the EoYS it was shown on a monitor in the gallery where ROVER was moving around.

3.2.5 EoYS 2013

ROVER was then shown in the 2013 Media Arts and Technology End of Year Show. While well received, several problematic issues arose. (see Table 3.2) The Melexis 90620, an infrared heat sensor array was used to find people in the space and worked surprisingly well. The data sent over the network was processed in Python, written to a database and visualized in Processing. The processing visualization was projected on the wall of the space. However, using the GoPro with a wifi backpack which streamed images to a laptop caused several seconds of latency which got progressively worse as the camera's SD card filled up. The laptop sent information to the Arduino via the xBee about where people were and what song to play, but the latency problem caused ROVER to sing regularly to

Table 3.2: EoYS 2013 Results

Expected	Actual	Cause
When ROVER sees a person it plays a song	ROVER plays a song after a while of viewing a person	Reading image off of GoPro via wifi took too much time
ROVER can detect people	ROVER could not detect people	Low lighting and wobbling camera caused blurry pictures
ROVER avoids obstacles	ROVER closelined by tables	Roomba bump sensors detect obstacles at floor levels not waist height
Everyone likes ROVER	ROVER terrified 2 children	Taller than kids, looks like a deformed muppet, unpredictable movement

where people used to be. Another problem with ROVER was that it had originally been tested in a well lit lab. When ROVER was shown in a poorly lit hallway, the camera used a slower shutter speed causing the images to be blurry. This, paired with the unstable platform on which it was installed, caused ROVER to be able to recognize faces with difficulty. Also, while the Roomba kept ROVER from running into walls, it did not keep it from knocking itself over by trying to go under tables. Finally, ROVER's form was very intimidating to small children at 5ft high with 2 different sized eyes peering out. People really wanted to touch ROVER, so making a robot that responded to touch would be an interesting future research topic.

Table 3.3: ROVER 2.0 Plans

Goal	Issue Cause	Solution
When ROVER sees a person it plays a song	Reading image off of GoPro via wifi took too much time	Camera module connected directly to raspberry pi
ROVER can detect people	Low lighting and wobbling camera caused blurry pictures	Camera module and more structural stability
ROVER avoids obstacles	Roomba bump sensors detect obstacles at floor levels not waist height	Proximity sensors at waist height
ROVER makes people happy	Taller than kids, looks like a deformed muppet, unpredictable movement	New design that was more structurally stable

3.3 ROVER 2 (Design)

After learning from the 2013 End of Year show, it was decided to continue with the project but the design of ROVER had to change to handle the problems with the previous design. (see Table 3.3) The Raspberry Pi Camera module came out in May 2013, so the issues with video quality and latency were solved by using a Raspberry Pi and the camera module instead of the laptop, GoPro with wifi backpack, and xBee. There was still a little latency, but the viewer would make the assumption due to ROVER's actions and sound as ROVER recognizing them as a person. Also the camera module did not have an issue with lower light environments. Four proximity sensors were added at 3 feet to keep ROVER from trying to go under tables. The structure was also changed to make it more stable. For the show it was decided to use the same audio generation algorithm.

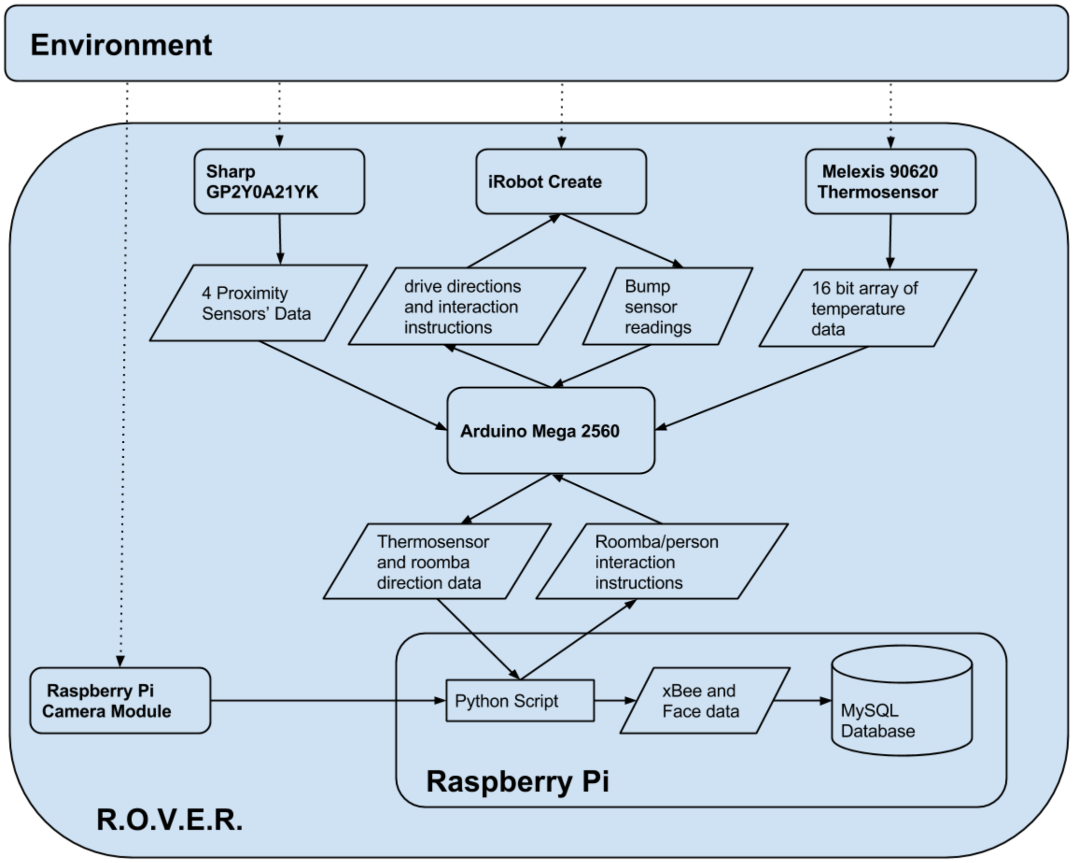

Environment
Sharp GP2Y0A21YK
iRobot Create
Melexis 90620 Thermosensor
4 Proximity Sensors' Data
drive directions and interaction instructions
Bump sensor readings
16 bit array of temperature data
Arduino Mega 2560
Thermosensor and roomba direction data
Roomba/person interaction instructions
Raspberry Pi Camera Module
Python Script
xBee and Face data
MySQL Database
Raspberry Pi
R.O.V.E.R.

Figure 3.7: ROVER 2.0 technology stack

3.3.1 Technology Stack

Switching to the Raspberry Pi allowed for a simpler design. Now all the processing could be done live onboard ROVER. To keep ROVER from being knocked over by going under tables or running into viewers, proximity sensors

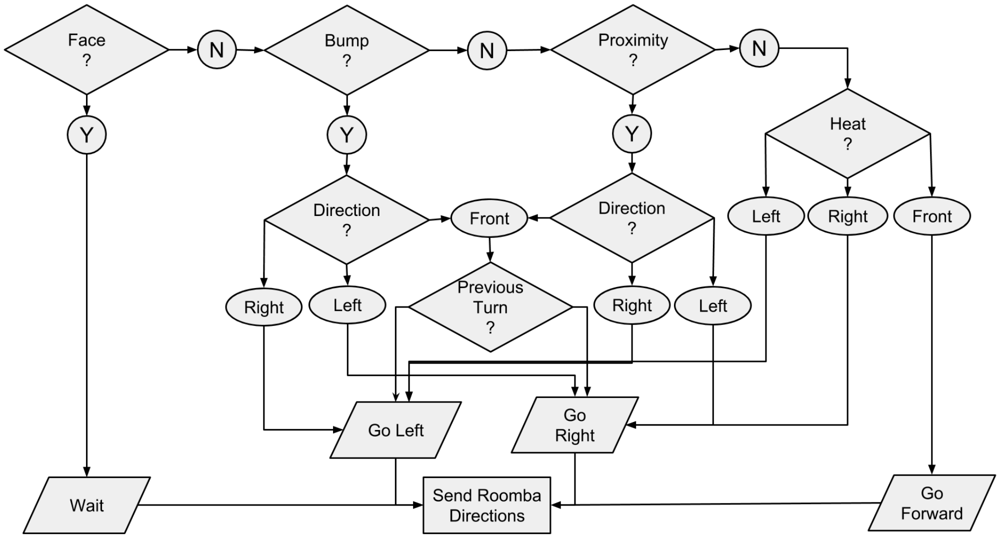


Figure 3.8: ROVER 2.0 movement algorithm

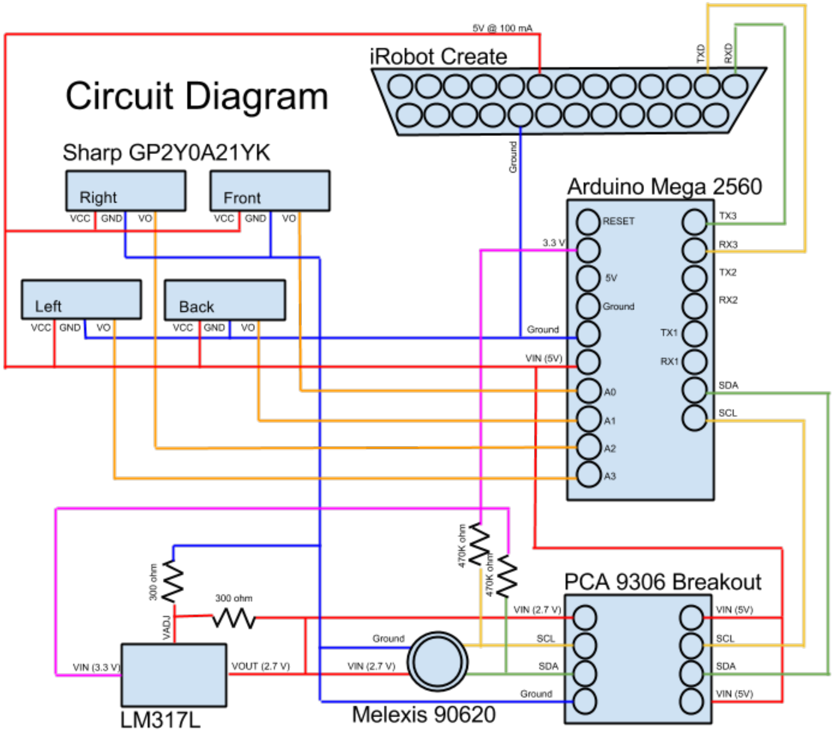


Figure 3.9: Wiring diagram for ROVER 2.0

were added. This affected the movement diagram, (see Figure 3.8) technology stack (see Figure 3.7) and wiring. (see Figure 3.9)

3.3.2 Design

A new structure that was more stable was designed and laser cut. (see Figure 3.10) The structure was 6 feet tall and made out of white acrylic. To be light, it was designed as a frame with one central spine, to which the cables were wired, and a series of ribs was added to create a form. At 5.5 feet there was a platform designed to house the electronics.

3.3.3 EoYS 2014

For the 2014 Media Arts and Technology Program End of Year show, ROVER 2.0 was presented. (see Table 3.4 for more details) The latency issues were fixed. The Raspberry Pi Camera worked better in low light, so there were not issues with blurry photos, and the new construction was more stable. By having ROVER's camera at a higher height, ROVER was better at detecting people's faces. However, at the show there were issues with powering the project. When it started to run low on power, the Roomba stopped responding to bump sensors. Also the song was trivial, so the next step was to research audio expression. People responded positively to ROVER, though the power issue meant that ROVER did

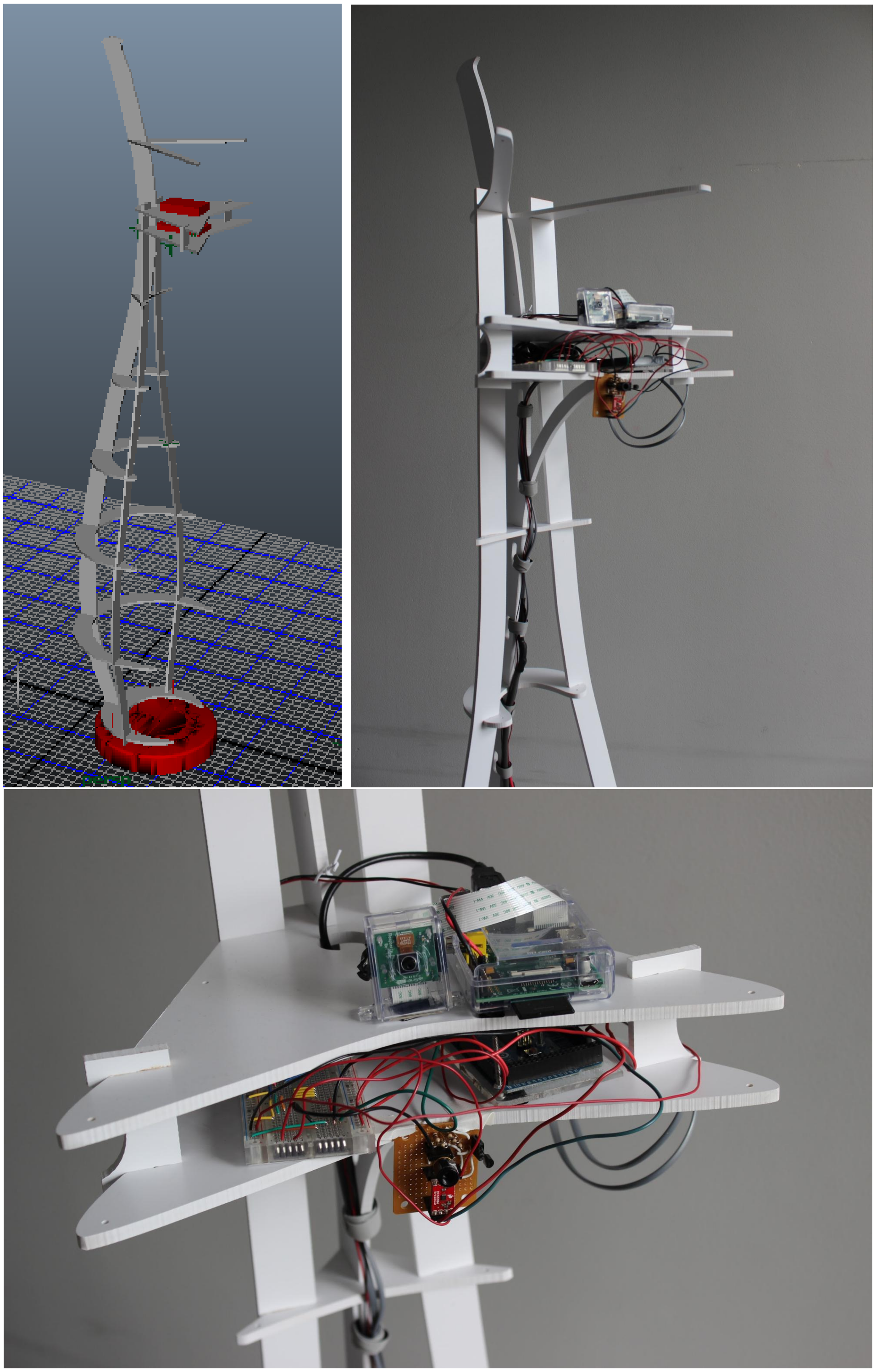


Figure 3.10: Design and final execution of ROVER 2.0

Table 3.4: EoYS 2014 Results

Goal	Solution	Actual
When ROVER sees a person it plays a song	Camera module connected directly to Raspberry Pi	worked
ROVER can detect people	Camera module and more structural stability	worked
ROVER avoids obstacles	Proximity sensors at waist height	Proximity sensors drew too much power. When running low, Roomba stopped responding to direction
ROVER makes people happy	New design was more structurally stable	Did not terrify children
ROVER finds song that makes people happy	Genetic algorithm choosing order of 5 notes	Did not converge on any sound

not get to interact with as many people as hoped. When ROVER did not stop even after seeing a person, some viewers tried to hug ROVER.

3.4 ROVER 3 (Sound Generation)

The 3rd version of ROVER focused on the sound generation algorithm and solving the power consumption issues.

3.4.1 Sound Generation Algorithm

The sound generation algorithm used is the first generative system for Non-Linguistic Utterances (NLUs) that uses both research from linguistics and music. The task of making melodies is generative, so the model was parametrized by

Table 3.5: Variables for Sound Generation Algorithm

Audio Aspect	Descriptor
Fundamental Frequency	tonic note of the key major or minor key the contour of the phrase (start) the contour of the phrase (end) variation around the contour
Amplitude	attack and sustain amplitude difference steady state amplitude mean steady state amplitude variance first contour value last contour value
Timbre	length of the attack decay sustain mean sustain variance release
Motive	motive length center motive length range pausing length mean pausing length variance

looking at the kinds of factors that people use to analyze sound. The 4 aspects of sound that were modified are fundamental frequency (F0), amplitude, articulation/timbre (envelope of the note), and motive. The function that describes each of these aspects takes 4-5 descriptors for a given musical phrase. (see Table 3.5 for more details)

Frequency

In a series of notes, the frequency of each note is dependent on 5 values: tonic note of the key, whether the song is in a major or minor key, the contour of the

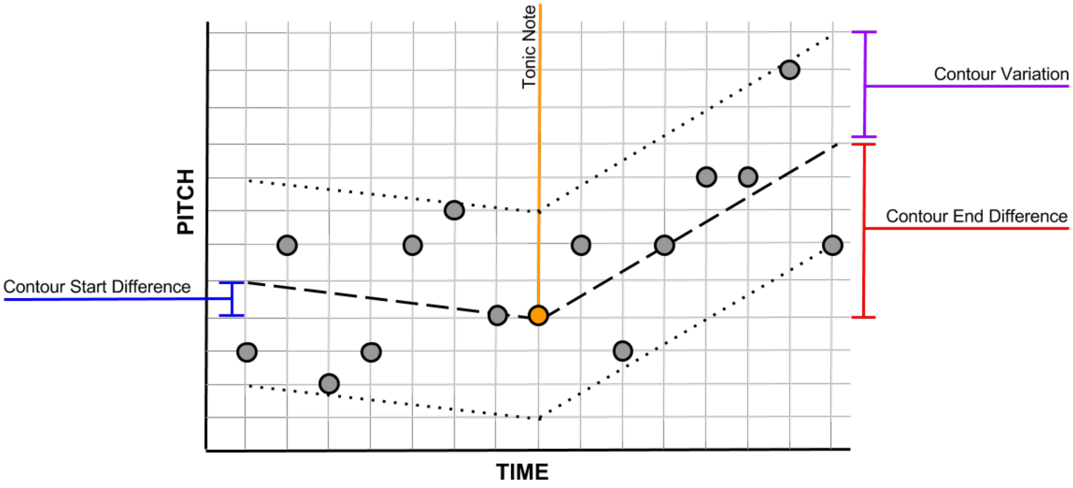


Figure 3.11: Melody of phrase algorithm

phrase (start and end values) and variation around the contour. (see Figure 3.11) First the tonic note and whether it is a major or minor key determines the scale. Then the contour, the rise and fall of the melodic line, is created by concatenating two linearly spaced vectors from the first contour value to zero, and from zero to the last contour value. This allows the pitch to ascend, descend and plateau at the beginning or end of the melody. The variation is a value that represents how much higher and lower the notes vary from the contour. Next each note's frequency is chosen by taking the tonic note and adding the random variation and contour value.

Amplitude

The amplitude of the attack and sustain of each note is dependent on 5 values: attack and sustain amplitude difference, steady state amplitude mean, steady state amplitude variance, first contour value and last contour value. First the contour is created by concatenating two linearly spaced vectors from the first contour value to zero, and from zero to the last contour value. The amplitude for each note is calculated from the steady state amplitude mean and variance which are used as variables for normal variate function to create a distribution of values. The contour is added to the steady state amplitude for the final amplitude. The attack amplitude is calculated by adding the attack and sustain difference to the sustain amplitude of each note.

Timbre

Timbre is created through an attack decay sustain release (ADSR) envelope. (see Figure 3.12) It is described by the length of the attack, decay, sustain mean, sustain variance and release. All of these values are described in milliseconds. First the sustain length is calculated from the mean and variance using a normal variate function to create a distribution of values. Next, using the attack and sustain amplitude, an ADSR envelope is created for each note. This is done by concatenating a series of linearly spaced vectors based on articulation lengths

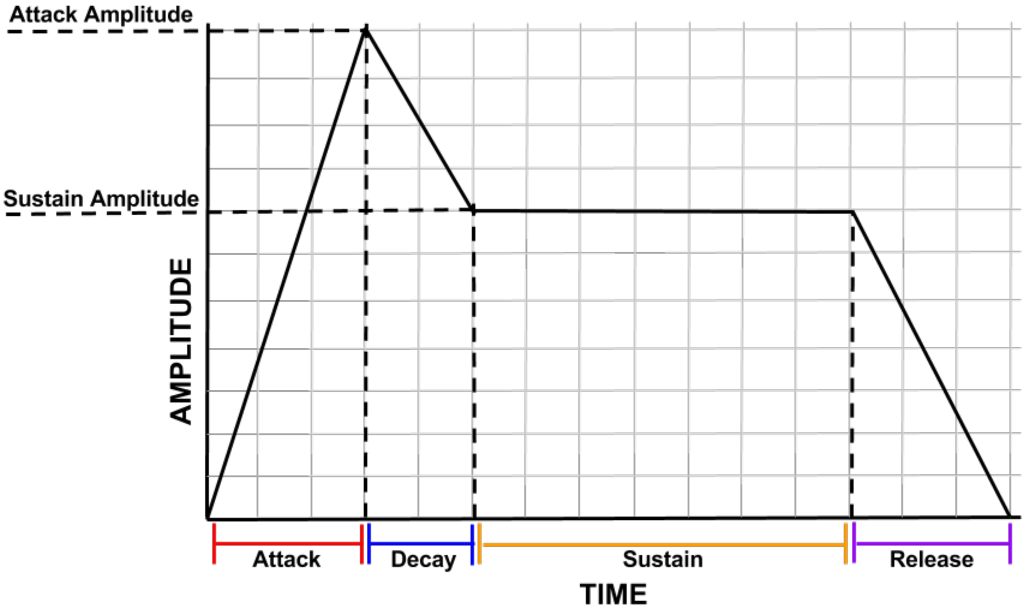


Figure 3.12: ADSR Envelope

from zero to attack amplitude, from attack amplitude to sustain amplitude, from sustain amplitude to sustain amplitude, and from sustain amplitude to zero.

Motive

The motive is described by the motive length center and range, as well as the pausing length mean and variance. (see Figure 3.13) The motive length describes after how many notes there is a pause. The motive length is calculated by choosing a random value within the motive range, until the sum of the motive lengths is greater than the length of the phrase. The length of each pause is calculated from

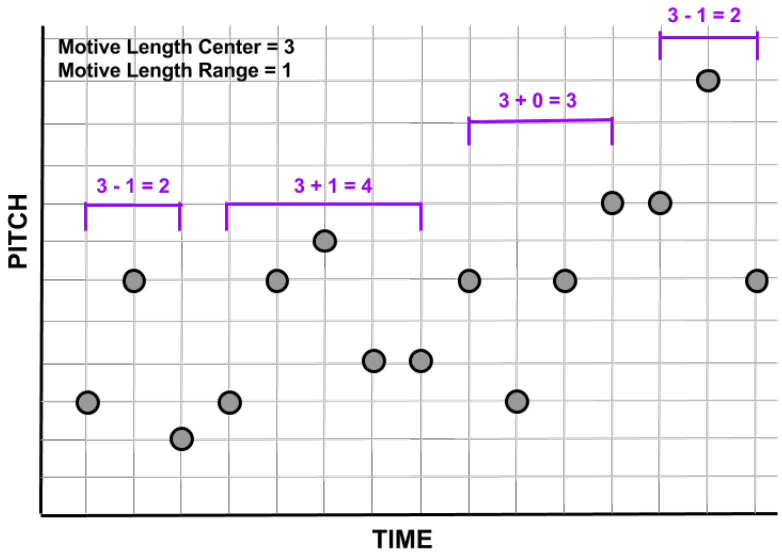


Figure 3.13: Motive length and range algorithm

the pause length mean and variance which are used as variables for normal variate function to create a distribution of values. The length is described in milliseconds.

Genetic Algorithm Mapping

Each parameter was mapped to a value between 0 and 1. The first generation of 50 was created by randomly assigning a value between 0 and 1 for each value, for each gene. Along with creating children via crossover and mutation, in each generation, 5 random new genes were added to the population. (see Table 3.6)

Table 3.6: Detailed parameters for sound generation algorithm

Audio Aspect	Descriptor	Type	Range	Starting Values
Fundamental Frequency	tonic note	int	3 to 103	3 to 103
	major/minor	int	0 to 24	0,4,8,12,16,20,24
	start contour	bool	0 or 1	0,1
	end contour	int	-40 to 40	-40,-20,0,20,40
	variation	int	-40 to 40	-40,-20,0,20,40
Amplitude	attack and sustain amplitude difference	float	-.4 to .4	-.40,-.20,0,.20, .40
	steady state amplitude mean	float	0 to 1	0,.2,.4,.6,.8,1
	steady state amplitude variance	float	0 to .24	0,.04,.08,.12,.16, .20,.24
	start contour value	float	-.4 to .4	-.40,-.20,0,.20, .40
	end contour value	float	-.4 to .4	-.40,-.20,0,.20, .40
Timbre	length of the attack	float	0 to 1	0,.2,.4,.6,.8,1
	decay	float	0 to 1	0,.2,.4,.6,.8,1
	sustain mean	float	0 to 1	0,.2,.4,.6,.8,1
	sustain variance	float	0 to .4	0,.2,.4
	release	float	0 to 1	0,.2,.4,.6,.8,1
Motive	motive length center	float	0 to 1	0,.2,.4,.6,.8,1
	motive length range	float	0 to .5	0,.1,.2,.3,.4,.5
	pausing length mean	int	1 to 9	1,3,5,7,9
	pausing length variance	int	0 to 5	1,2,3,4,5

Table 3.7: EoYS 2015 Results

Goal	Solution	Actual
ROVER avoids obstacles	Separate power for proximity sensors	worked
ROVER finds song that makes people happy	Algorithm using vocal qualities	Bug in program caused it to crash whenever a song was in minor mode. Did not converge on any song, too many variables.

3.4.2 Technology Stack

In ROVER's final implementation, the Raspberry Pi produced the audio, which required the addition of speakers. The sound was created based on audio research and using the library PyAudio. PyAudio is not designed for realtime audio synthesis so the sounds had to be pre- generated in batches. Since the Raspberry Pi was producing the sound directly, there was a decrease in latency. The Raspberry Pi was connected to wifi so that SSH (Secure Shell) could be used to start the program and see diagnostics.

3.4.3 EoYS 2015

ROVER was presented in the 2015 EoYS. (see Table 3.7) The main difference between the previous year's project and the 2015 show was the sound generation algorithm. There was a bug in the sound generation algorithm which would crash the Python script on occasion, requiring the system to be rebooted. Since

ROVER was on wifi it was easy to restart. Later the problem was solved for the pilot study. The public responded positively to ROVER. ROVER was attracted to a light during the opening speech and moved toward the light interrupting the speech.

Chapter 4

Pilot Study

Since the public responses to ROVER at the End of Year Show were very encouraging, a more controlled study of human robot interaction using ROVER was the next step. The Institutional Review Board required the participants' signed consent for the videos to be saved for further analysis. To get viewers' permission to have their video recorded was impossible in a public space. Therefore the data collected could not be used for research. The purpose of the study was to collect data to see how different audio qualities of sound coming from different sources affect a viewer's emotional response, specifically looking at the participant's response to sounds coming from a computer, an immobile robot and a mobile robot. It was theorized that mobility and the robot's approach would caused a stronger emotional response in the viewer. The results would be used to further human computer interaction by creating a means for robots to be emotionally expressive.

The results of prior research on emotional response to specific audio aspects were tested, using different audio qualities to provoke emotion. Two emotional states were selected, sad and excited. Sad is defined as low valence, low arousal and submissive. Excited is defined as high valence, high arousal and dominant. A sad emotional state is expressed in human speech by low base frequency, small frequency range, low speech rate and high pause rate. An excited emotional state is expressed in human speech by high base frequency, large frequency range, high speech rate and low pause rate. The sound envelope was varied to affect arousal by a percussive envelope or a steady state envelope. Two musical modes were tested which affected valence, major and minor mode.

Twenty subjects were requested to participate for 30 minutes, each hearing 24 sounds. The data that was recorded was what stimuli was used, the video of the participants' reactions, the survey responses and an anonymous participant number. The user participated in 3 different phases of the study in a varied order. Each phase took 7-10 minutes during which the participant listened to eight 5-second sounds. After a participant was recorded while listening to a sound, they filled out a Self Assessment Manikin (SAM) survey.(Morris 1995) The facial expression data was then analyzed and compared to the Self Assessment Manikin survey results. A pilot study was run to solidify the procedure.

The blocks were: participant interacting with the computer, participant approaching ROVER and interacting with it, and ROVER approaching the participant and interacting with her. When the participant approached ROVER, she was initially asked to stand 13 feet from ROVER and then to walk up to ROVER. When she got close enough to ROVER that her face was at least 10 pixels wide in a 320X240 pixel image, ROVER started playing the first song. When ROVER approached the participant, the participant was asked to stand 10 feet from ROVER, the investigator pressed a button on ROVER, and ROVER moved toward the participant until the participant's face was 10 pixels wide. There it stopped and played the first melody. For the computer block, the participant listened to the sounds and responded to the survey on the computer. A webcam was used to record the interaction with the computer. After each block the participant filled out a questionnaire to determine the level of embodiment portrayed.

4.1 Technology Stack

The user study required a server to host the website for collecting data. The server ran Ubuntu. A website was built using Django as a front end and MySQL as a back end. The database design is described in Figure 4.1. The sound files

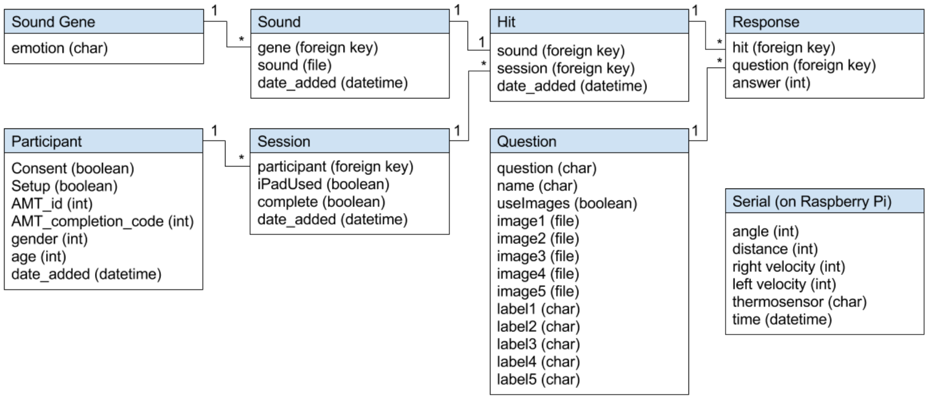


Figure 4.1: Database Design

were pre-generated and hosted on both the server and the Raspberry Pi. (see Figure 4.2)

4.2 Sound Generation Algorithm

To be an emotive robot that can create an emotive expressive response in the viewer, a generative system to create these sounds needed to be designed. Ideally there would be a relationship between the final sounds and research in music and linguistics. To start, research was done on emotion and audio to see what types of audio qualities are expressive. After defining what audio qualities would be the study's focus, parameters and functions were defined for creating the sounds that would be generated. Two sets of parameters were designed, one that expressed sadness and the other that expressed excitement. The sound generation

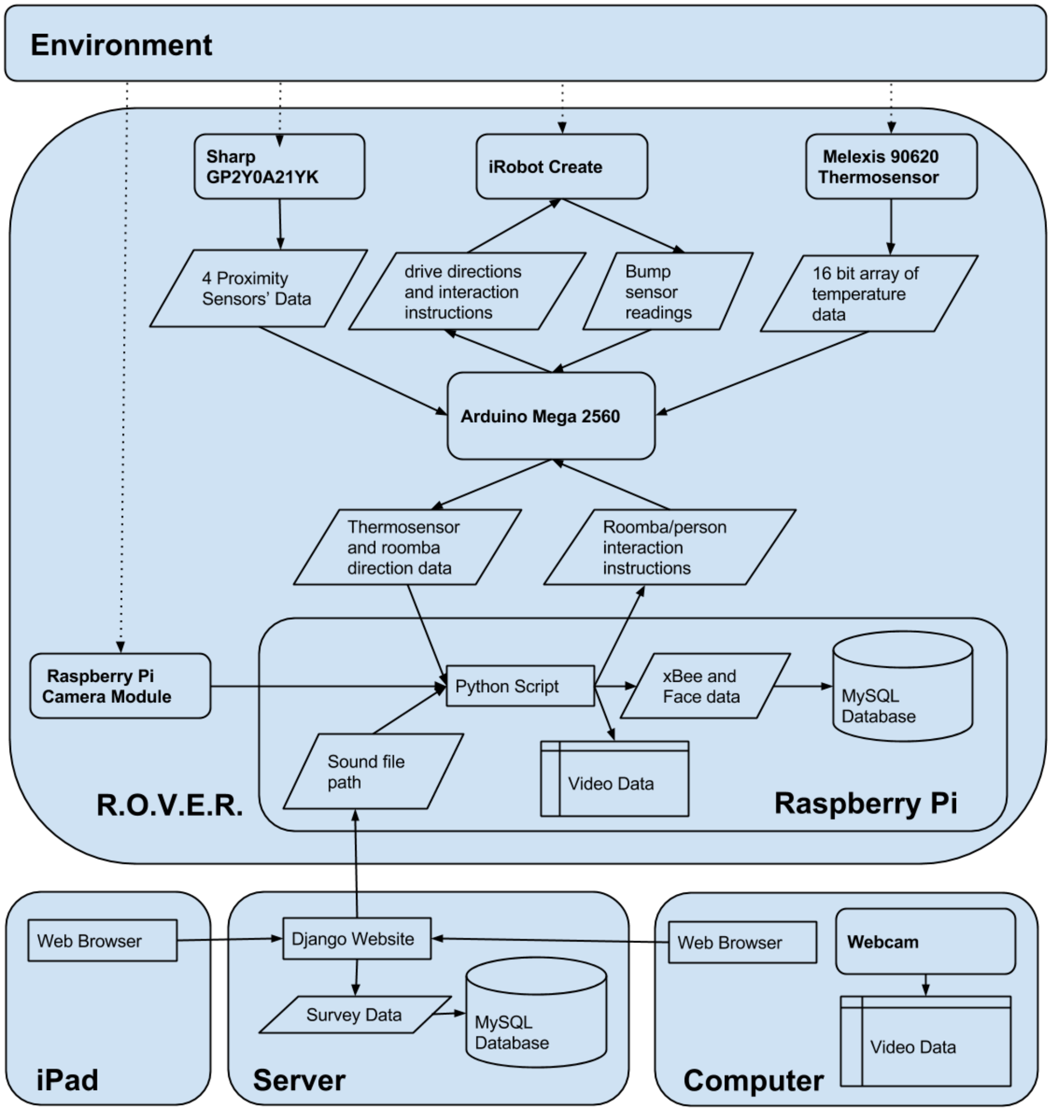


Figure 4.2: Pilot Study technology stack for ROVER

Table 4.1: Descriptors for two emotions

Emotion	“Sad”	“Happy”	Reference
Base F0	528.35	1500	(Read 2014)
F0 range	570.31	1460.77	(Read 2014)
Speech Rate	Low	High	(Scherer et al. 2003)
Pause Ratio	High	Low	(Scherer et al. 2003)
Envelope	Steady State	Percussive	(Huron et al. 2014)
Pitch Contour	Rising	Flat	(Allan 1984)
Mode	Minor	Major	(Turner & Huron 2008)

algorithm used the system described in 4.4.1, but only created two kinds of sounds, the stereotypically “happy” and “sad” sounds. The parameters it used were base fundamental frequency, frequency range, speech rate, pause ratio, envelope, pitch contour and mode. (see Table 4.1)

4.3 ROVER Movement Algorithm

The new ROVER Movement Algorithm was similar to the original movement algorithm (4.2.1 Movement Algorithm Diagram) but in this case, if a sound hadn’t been sent, ROVER waited a second and checked again. If a sound had been sent and a face was seen, then the movement paused until the Roomba was power-cycled.

ROVER’s code had 3 threads, a sound thread, a video thread and a serial thread. (see Figure 4.3) The sound thread would check the flag soundToPlay. If it was false, it would go to the server to see if there was a sound that was

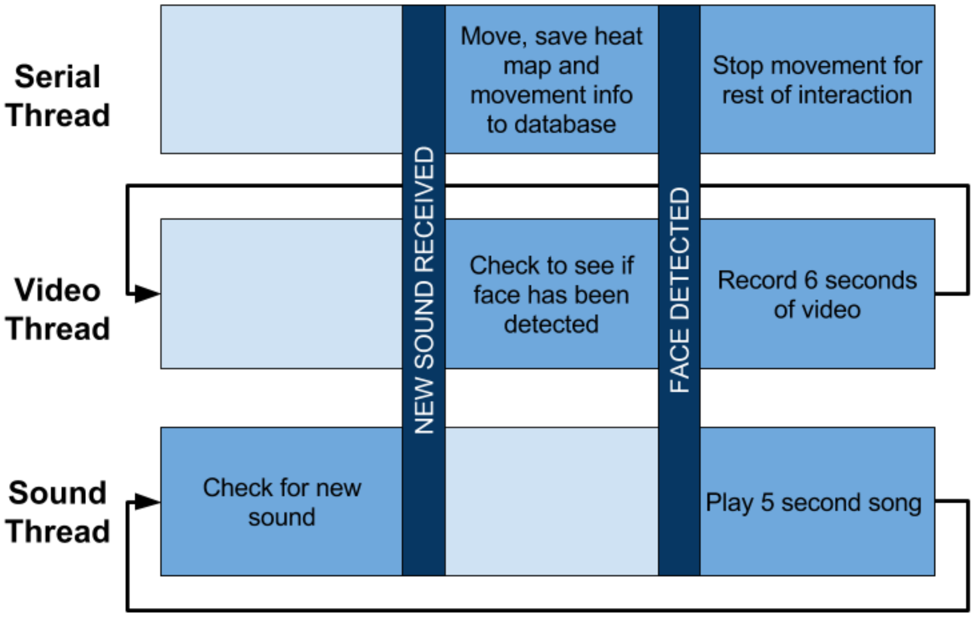
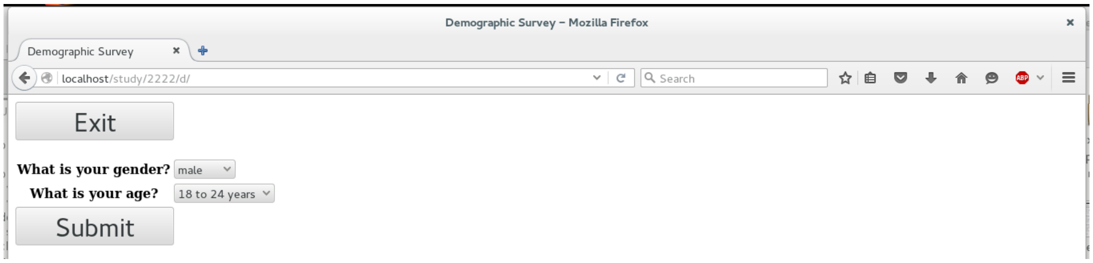

Serial Thread
Video Thread
Sound Thread
NEW SOUND RECEIVED
FACE DETECTED
Move, save heat map and movement info to database
Stop movement for rest of interaction
Check to see if face has been detected
Record 6 seconds of video
Check for new sound
Play 5 second song

Figure 4.3: Pilot Study program design

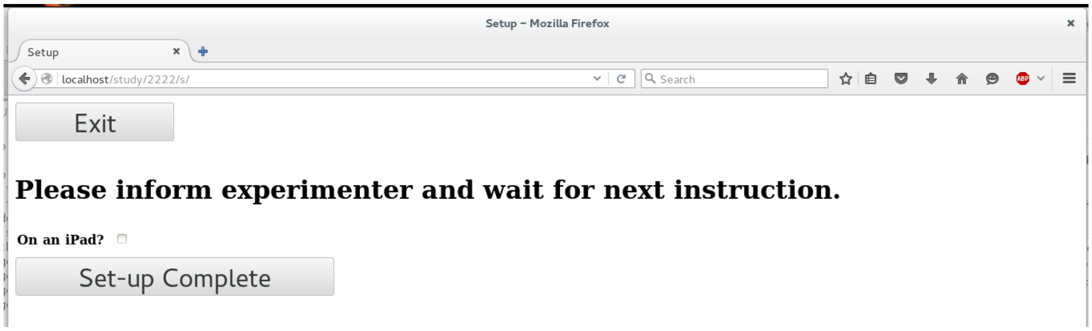
different from the last sound played, and if there was a new sound, then it would set soundToPlay equal to true. If soundToPlay equaled true, and the faceVisible flag was true, it would play the sound. The video thread waited till soundToPlay was true, and then it started checking the video for faces. Once it saw a face, it would record a 5 second larger format video. In the serial thread if the face was visible, then it sent a message to ROVER to wait forever, but if there was no sound to play and no face is visible, then it sent a message to ROVER to wait a second. If there was a sound to play but a face was not seen yet, the serial thread read serial data and recorded it in the database.



What is your gender? male

What is your age? 18 to 24 years

Figure 4.4: Demographic survey user interface



Please inform experimenter and wait for next instruction.

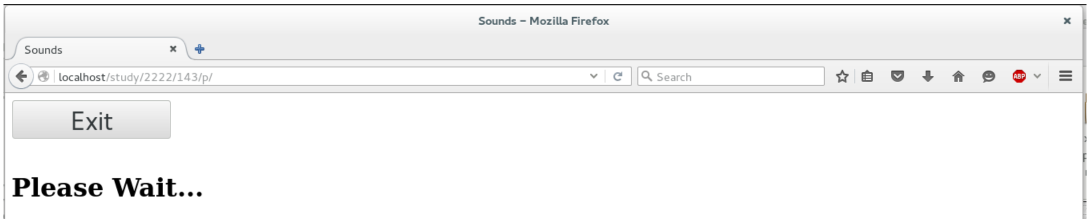
On an iPad? □

Figure 4.5: Setup page user interface

4.4 Web Interface Design

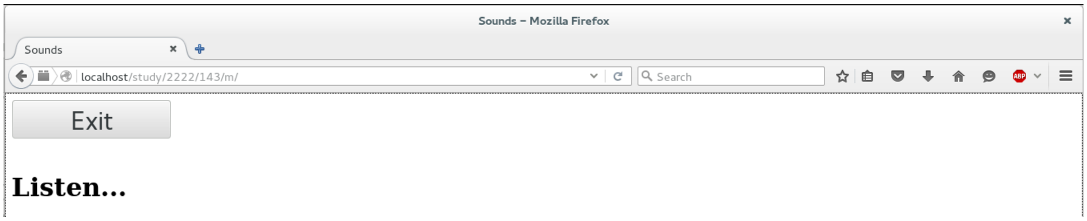
The web interface has 7 pages. The first page “start” takes an amtId and sends the user to the demographics page, if the participant “agrees.” Since the participant signed a consent form, the proctor presses the “I agree” button after they sign the form. (see Figure 4.4)

The demographics page takes an amtId and prompts the user with a short demographic survey. Once completed, the computer creates a participant database entry with demographic results.



Please Wait...

Figure 4.6: Wait page



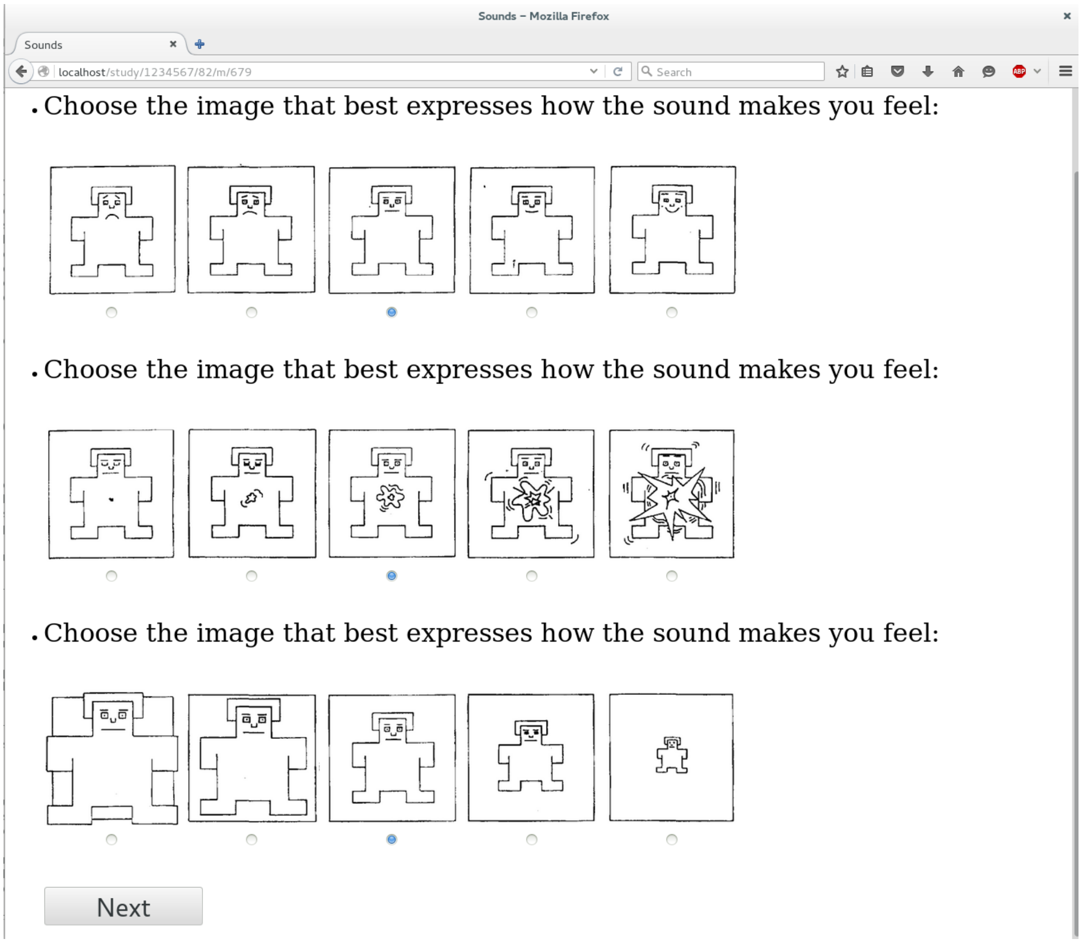
Listen...

Figure 4.7: Listen page

The participant is then directed to the setup page which takes an amtId. The setup page has the proctor input whether the iPad is used. Once complete, the page creates a session entry with whether or not the iPad was used. (see Figure 4.5)

The wait page only appears if the iPad is not used and forces the participant to wait 6 seconds between stimuli. The wait page takes an amtId and sessionId. (see Figure 4.6)

The listen page takes an amtId and sessionId. (see Figure 4.7) It gets sounds that no one has listened to and checks to see how many sounds the participant has heard in this session. If the participant has listened to 8 or more songs, it



- Choose the image that best expresses how the sound makes you feel:

○ ○ ● ○ ○

- Choose the image that best expresses how the sound makes you feel:

○ ○ ● ○ ○

- Choose the image that best expresses how the sound makes you feel:

○ ○ ● ○ ○

Next

Figure 4.8: Survery response user interface

saves the session as complete and redirects to the setup page. If the participant has heard fewer than 8 songs, it checks to see if they have heard 4 of one sound type. If so, it filters the unheard sounds by sound type, then chooses a random sound from the unheard sounds, creates a hit for that sound, and plays the sound in the browser of the computer.

Table 4.2: Emotional response survey descriptors and mapping

Scale	-2	0	2
Pleasure	Unhappy Unsatisfied Annoyed	Neutral	Happy Satisfied Pleased
Arousal	Relaxed Sleepy Calm	Neutral	Stimulated Awake Excited
Dominance	Controlled Small Influenced	Neutral	Controlling Big Influential

The soundName page is accessed by the Raspberry Pi and shows the sound files path of the most recent hit created within the last 5 seconds.

Once the sound is finished playing, the questions page appears which takes an amtId, soundId and sessionId. The question page shows questions and saves the responses to questions in a database associated with the hit. (see Figure 4.8) Each question was explained to the participants with 6 descriptors for each scale. (see Table 4.2) (Morris 1995)

Chapter 5

Results

5.1 Analysis

Each question from the SAM results was analyzed using two-way repeated measures ANOVA with the program SPSS Statistics. While there was not enough data to have any interactions be statistically significant, within the sound and block results there were some statistically significant results. The scale for each question was a -2 to 2 scale, with 0 as neutral. (ref. table in 4.4)

The difference in valence between the two sound types was significant with a p value of 0.00008. (see Figure 5.1) On average the participant rated the "happy" sounds as a 0.923 (std dev. 0.154), while the "sad" sounds were rated as a -0.208 (std dev 0.172). The participants responding in the survey rated their emotional response to the "happy" sounds, happier, and the "sad" sounds, sadder.

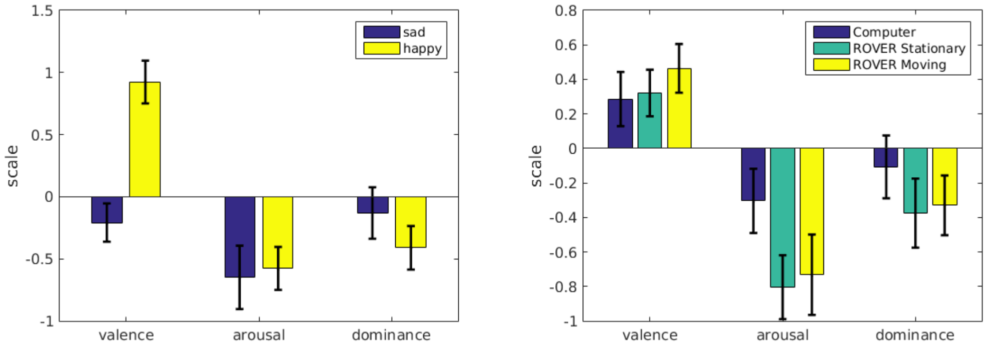


Figure 5.1: Emotional response to "sad" and "happy" songs. (left) Emotional response in different blocks. (right)

The difference in arousal between blocks was significant with a p value of 0.005. (see Figure 5.1) There was no significant difference between ROVER stationary and moving, but there was a significant difference between the computer block and the two different ROVER states. The average arousal for the computer, ROVER stationary and ROVER moving was -0.304 (std. dev. 0.186), -0.804 (std. dev. 0.185), and -0.732 (std. dev. 0.233) respectively. This shows that the participants felt more aroused when interacting with the computer than with ROVER. There was not a significant difference between interacting with a stationary or moving ROVER. These results were surprising and could be explained by proximity to the sound source or volume level of the sound. The difference in dominance between blocks was almost significant with a p value of 0.089. This could be explained by

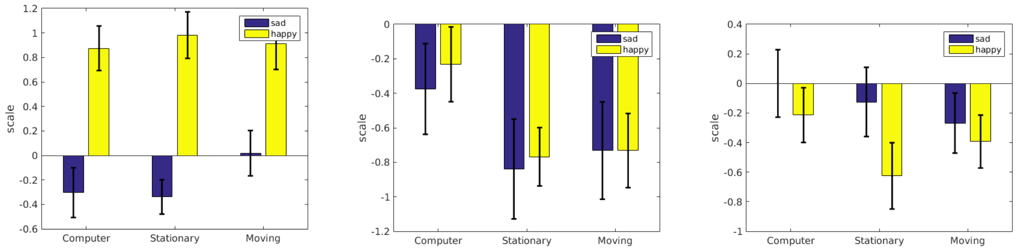


Figure 5.2: Song VS Block, for (left to right) Valence, Arousal, Dominance

the difference in response to happy sounds between the computer block and the ROVER stationary block.

Since there was not enough data, the relationship between block and sound was not statistically significant. (see Figure 5.2) For the valence question, participants found the sad sound from the moving robot less sad than when it came from the stationary robot. For the arousal question, there was little deviation between sad and happy sounds within blocks. For the dominance/submission question, in the block during which the participants interacted with stationary ROVER, participants found the happy sounds caused a stronger feeling of submission. Also in reaction to the "happy" sound, when compared to the computer, ROVER stationary caused a stronger feeling of submission.

5.2 Discussion

There were many problematic issues with the study. The participants were not properly trained, so participants looked at the iPad instead of at ROVER, which influenced video quality. Participants would often hit the next button without hearing the sound. The speakers and volume levels were different for ROVER and the computer. ROVER's movement was inconsistent and sometimes it would move toward the server or the door. ROVER's movement from one participant to the next was not consistent due to power issues. Frequently the prompter had to restart the movement phase. Once the battery was replaced, the movement block became more consistent. ROVER stopped at different distances from different participants, depending on whether or not ROVER could detect the participant's face. This occurred because participants were not always looking at ROVER. There was also a difference in distance between ROVER and the participant, and the computer and the participant. Also the participant had to sit to interact with the computer and input the survey results directly into the computer, while the participant had to stand to interact with ROVER, while inputting the survey information into an iPad. Finally there was a non-standard script for explaining the survey questions. This led to some confusion about the submission/dominance spectrum.

Due to the fact that participants did not look at ROVER, the video data was inconsistent. With the computer, participants looked at the webcam, but they had to look down for the ROVER interactions which made the participant's face less visible and not conducive to emotion detection techniques. This could be solved by having a training phase for participants where they would be trained to look at ROVER and not hit the next button until they heard a sound.

The higher arousal for interacting with the computer compared to ROVER could be due to multiple factors that were not specific to ROVER. It could be caused by the way the survey data was taken, by interacting with an iPad vs the computer directly. It could be caused by distance, since ROVER would stop 3 feet away from participants, but the participants were closer to the computer. Edward T. Hall's personal reaction bubbles would define the interaction with ROVER as in social space, while the interaction with the computer was within personal space. Personal space is described as the area surrounding a person which they view as their space. Most people feel anxiety, anger or discomfort when their personal space is encroached upon. (Hall 1966) This could cause anxiety or heightened arousal when interacting with the computer compared to ROVER. Also the computer had a higher volume level than ROVER. In prosody research, amplitude is correlated with increased arousal. For example, when someone is excited they speak louder. It would make sense then that a louder sound would make people excited. To make sure that these factors did not contribute to the difference in arousal, the final study would make sure that ROVER stopped in a consistent location. Also the participant would interact with the iPad when using the computer and stand the same distance from the computer as the participant did from ROVER. Finally the same speakers and volume level would be used for both the computer and ROVER.

Chapter 6

Future Directions and Conclusion

6.1 Potential Directions

There are many directions in which this project could continue. For example, different spaces and contexts could be explored. ROVER has been shown both in an academic study and in a gallery. It would be interesting to see if there is a difference in emotive response in a gallery with a single person, in a gallery with multiple people, in a study with one person or in a study with multiple people. Another direction would be to explore a wider range of emotions. Other emotional sounds could be created and tested using this system. To look at different physical forms, for the class morphogenesis, a project was done to explore the form and embodiment for ROVER, specifically to creating a skin for ROVER. (see Figure 6.1 & 6.2) The surface of the project was created and modified using the Catmull-Clark algorithm, (Catmull & Clark 1978) spherical harmonics

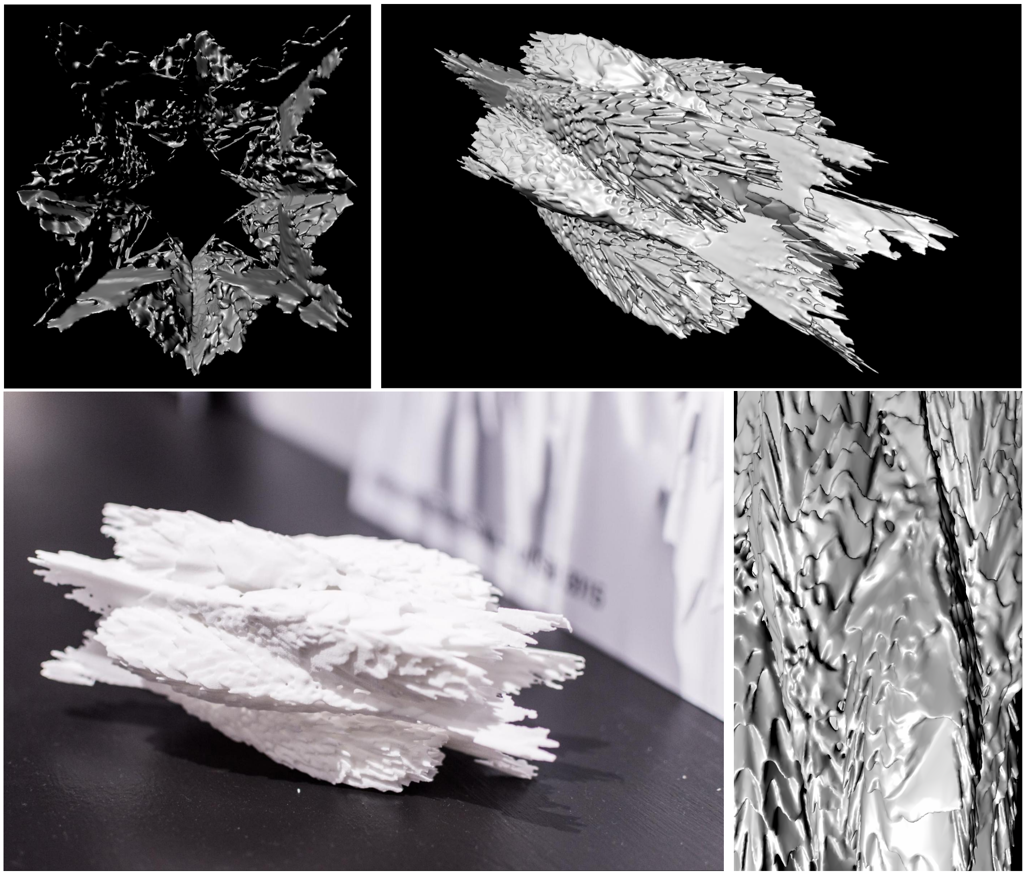


Figure 6.1: Renderings and final product from morphogenesis class. Photo credit: Mohit Hingorani

and Michael Hansmeyer's "Design by Subdivision" algorithm. (Hansmeyer et al. 2010)

As always there are technical changes which could improve the project, particularly in the movement algorithm. A proportional-integral-derivative (PID) controller could be used to smooth the movement. ROVER could be programmed to remember where people were previously located and go back to more densely

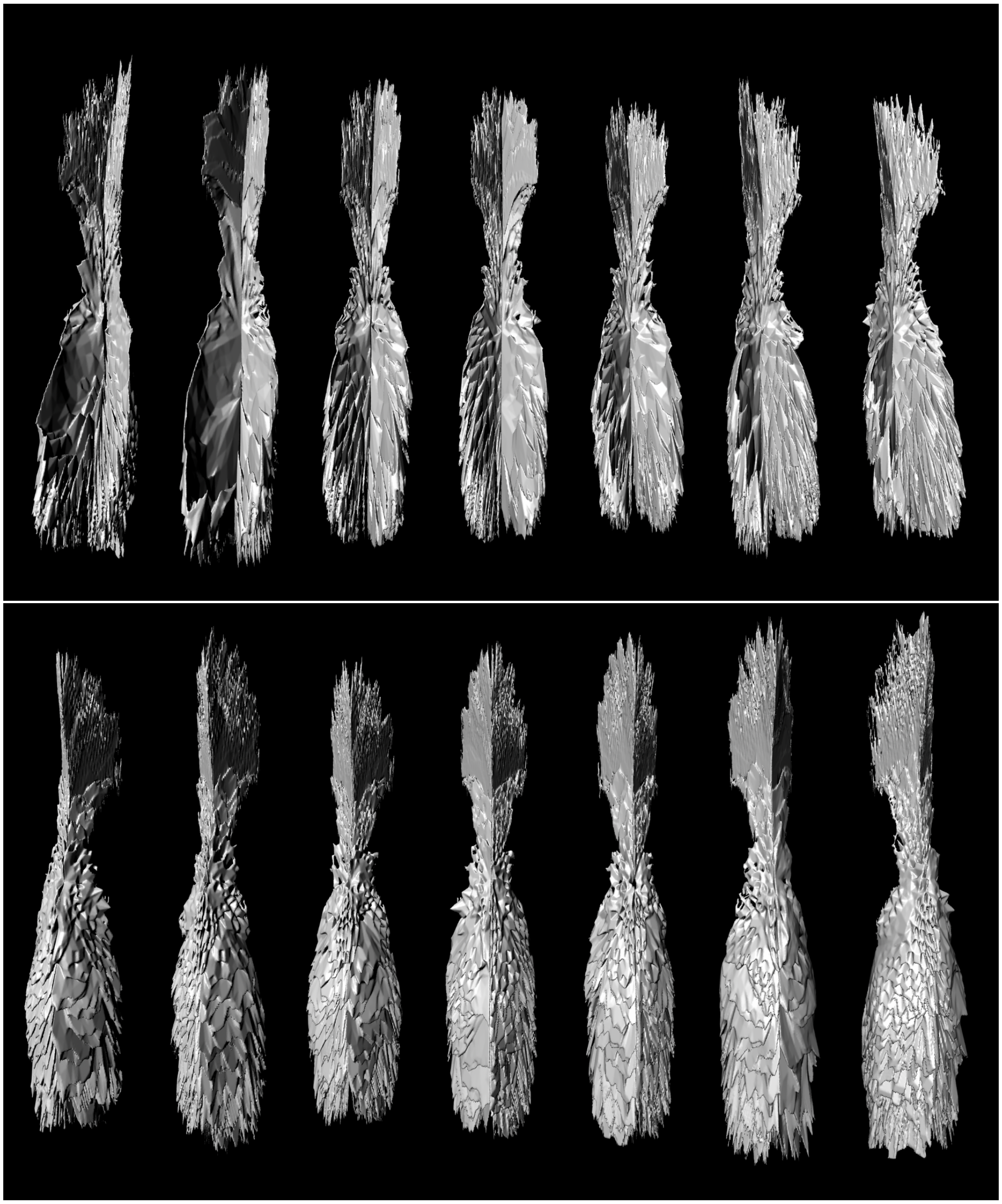


Figure 6.2: Renderings of different skins for morphogenesis class.

populated areas when alone. ROVER could also be programmed to look around first before moving to make sure it moves in the best direction, instead of just moving toward what it can see in it's immediate view.

6.2 Future Research

Further directions of research will involve levels of proxemics in both the digital and physical realms, looking at the emotional response to sound, focusing on discomfort and the uncanny valley. Three levels of engagement, embodiment and proximity will be studied, a physical robot in the social space, a digital robot in the social space and a physical robot in the participant's intimate space. Throughout the investigation of different levels on embodiment and proxemics, the focus would be on emotive sound for interaction, particularly unease created by the uncanny valley. This could be done by creating a dissonance between expected emotion and emotion expressed. An extension would be to look directly at the uncanny valley of speech, particularly studying the difference between speech, computer generated emotive speech, gibberish and non-linguistic utterances.

The pilot study was the beginning of research about a physical robot in the social space. The pilot study's results strengthen the precedence for a full study. The first and most direct extension is to run the full study with 40 participants and the changes described in the results section. This would allow for video data to analyze emotional reactions. Also the potentially confounding issues, like volume level and proximity could be controlled. This line of research will also be brought further into the physical through an interaction in intimate space, looking at emotive response to sounds produced in relationship to haptic engagement. This would be done with soft robotics, touch sensors and/or conductive fur.

The next step of the project is to extend it into the virtual realm. A study will be run to look at the difference in emotive response to a physical robot, a software robot on the computer, a software robot in the AlloSphere, a telepresent robot on the computer and a telepresent robot in the AlloSphere. For the telepresent robot, the Nokia Ozo, a spherical and stereoscopic video capture system with a 360x360 surround sound array could be used. Computer vision, surveys, and potentially the Four Eyes Lab's EEG and/or AlloSphere Research Group's biopack could be used for measuring the participants' responses. If there is emotional engagement with a software robot in the AlloSphere, then the AlloSphere would be an innovative way to prototype different physical forms and gestures of a robot without actually building the physical robot. This would cut down on prototyping time and materials, and allow for rapid analysis of form and gesture.

6.3 Conclusion

For this project a robot was designed and built for studying human-robot interaction. It was shown at four art exhibitions and used as part of a pilot study on human-robot interaction. A system was defined and built based on prior work as a part of this project for creating emotive sound. To run the user study, a system was built for users to self report emotive response to stimuli. Finally a pilot study was run that tested participants' responses to two different types of sounds in 3 different situations. Results confirmed that the sounds created expected emotive responses, specifically that "happy" sounds increased valence significantly. The results disproved the hypothesis that the robot created a more emotionally aroused interaction. This could be due to proximity to the source or volume level of the sound. These factors will be adjusted for when re-running the study.

As robots become a part of the household and are used more in the service sector, it is important to create a seamless interaction with novice users. Humans treat technology like they treat people and expect it to respond like a person, including being able to emote vocally. The research conducted for this project is a proof of concept and foundation for further research in the field of emotive vocal communication in robotics. By algorithmically generating emotive responses, the technology becomes more lifelike and easier to interact with. The ROVER's sound generation algorithm could be used for emotive robots, smart devices and computer applications.

Bibliography

Ablinger, P. (2006), 'Quadraturen', http://ablinger.mur.at/docu11.html. Accessed: 2016-03-17.

Adi, P. G. (2008), 'Alexitimia: An autonomous robotic agent', *A Minima* (23).

Aldebaran (2015), 'Pepper, the humanoid robot from aldebaran, a genuine companion', https://www.ald.softbankrobotics.com/en/cool-robots/pepper. Accessed: 2016-06-24.

Allan, K. (1984), 'The component functions of the high rise terminal contour in australian declarative sentences', *Australian Journal of Linguistics* **4**(1), 19–32.

Apple, W., Streeter, L. A. & Krauss, R. M. (1979), 'Effects of pitch and speech rate on personal attributions.', *Journal of Personality and Social Psychology* **37**(5), 715.

ARCHIGUIDE (2000-2016), 'Vsba venturi scott brown associates', http://www.archi-guide.com/AR/venturi.htm. Accessed: 2016-06-30.

Asimov, I. (1950), *I, Robot*, Gnome Press.

Atzori, P. & Woolford, K. (2015), 'Extended-body: Interview with stelarc', *CTheory* pp. 9–6.

Badger, J. (2008), 'Reincribing rings', http://jeffbadger.com/art_2008_rerings.html. Accessed: 2016-03-17.

Bahadori, S., Cesta, A., Grisetti, G., Iocchi, L., Leone, R., Nardi, D., Oddi, A., Pecora, F. & Rasconi, R. (2003), Robocare: an integrated robotic system for the domestic care of the elderly, *in* 'Proceedings of Workshop on Ambient Intelligence AI* IA-03, Pisa, Italy', Citeseer.

Bakkum, D. J., Gamblen, P. M., Ben-Ary, G., Chao, Z. C. & Potter, S. M. (2007), 'Meart: the semi-living artist', *Frontiers in neurorobotics* **1**, 5.

Banse, R. & Scherer, K. R. (1996), 'Acoustic profiles in vocal emotion expression.', *Journal of personality and social psychology* **70**(3), 614.

Bartneck, C. (2000), 'Affective expressions of machines', *emotion* **8**, 489–502.

Bartneck, C. (2002), *EMuu: An Embodied Emotional Character for the Ambient Intelligent Home*, Technische Universiteit Eindhoven.

Bekey, G. & Yuh, J. (2008), 'The status of robotics', *Robotics & Automation Magazine, IEEE* **15**(1), 80–86.

Bell, T. (1985), 'Robots in the home: Promises, promises: While great expectations are held for certain robot types, the robots for fun and educational purposes are limited in their adaptability to useful tasks', *Spectrum, IEEE* **22**(5), 51–55.

Benthall, J. (1972), *Science and technology in art today*, Thames and Hudson.

Berk, J. (2009), 'Alavs', `http://www.alavs.com/`. Accessed: 2016-03-17.

Bethel, C. L. & Murphy, R. R. (2006), Affective expression in appearance constrained robots, *in* 'Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction', ACM, pp. 327–328.

Bolygó, B. (2015), 'Bálint bolygó: Works'. Accessed: 2016-06-30.

Bourguet, M.-L. (2003), Designing and prototyping multimodal commands., *in* 'INTERACT', Vol. 3, Citeseer, pp. 717–720.

Breazeal, C. (2003), 'Emotion and sociable humanoid robots', *International Journal of Human-Computer Studies* **59**(1), 119–155.

Breazeal, C. L. (2004), *Designing sociable robots*, MIT press.

Breitenstein, C., Lancker, D. V. & Daum, I. (2001), 'The contribution of speech rate and pitch variation to the perception of vocal emotions in a german and an american sample', *Cognition & Emotion* **15**(1), 57–79.

Brown, P. R. (1996), The influence of amateur radio on the development of the commercial market for quartz piezoelectric resonators in the united states, *in* 'Frequency Control Symposium, 1996. 50th., Proceedings of the 1996 IEEE International.', IEEE, pp. 58–65.

Campbell, D. T. (1958), 'Common fate, similarity, and other indices of the status of aggregates of persons as social entities', *Behavioral science* **3**(1), 14–25.

Catmull, E. & Clark, J. (1978), 'Recursively generated b-spline surfaces on arbitrary topological meshes', *Computer-aided design* **10**(6), 350–355.

Cod.Act (2011), 'Pendulum choir', `http://codact.ch/gb/pendugb.html`. Accessed: 2016-03-17.

Cohen, H. (1995), 'The further exploits of aaron, painter', *Stanford Humanities Review* **4**(2), 141–158.

Corbett, C. et al. (2012), 'Live from mars', *Monthly, The* (Sept 2012), 14.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J. G. (2001), 'Emotion recognition in human-computer interaction', *Signal Processing Magazine, IEEE* **18**(1), 32–80.

Criqui, J. (2001), 'Eater's digest (wim delvoye's' cloaca', 2000)', *ARTFORUM* **40**(1), 182–183.

Dalla Bella, S., Peretz, I., Rousseau, L. & Gosselin, N. (2001), 'A developmental study of the affective value of tempo and mode in music', *Cognition* **80**(3), B1–B10.

Darwin, C., Ekman, P. & Prodger, P. (1998), *The expression of the emotions in man and animals*, Oxford University Press, USA.

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M. L., Koay, K. L. & Werry, I. (2005), What is a robot companion-friend, assistant or butler?, *in* 'Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on', IEEE, pp. 1192–1197.

Davidson, R. J., Scherer, K. R. & Goldsmith, H. (2003), *Handbook of affective sciences.*, Oxford University Press.

Debatty, R. (2009), 'Self portrait machine', `http://we-make-money-not-art.com/selfportrait_machine/`. Accessed: 2016-03-17.

Demers, L.-P. (2015), 'Machine performers: Agents in a multiple ontological state'.

DeRosia, B. (2014), 'Brian derosia', `http://www.brianderosia.com/`. Accessed: 2014-03-17.

Dudley, H., Riesz, R. & Watkins, S. (1939), 'A synthetic speaker', *Journal of the Franklin Institute* **227**(6), 739–764.

Ekman, P. & Friesen, W. V. (1969), 'The repertoire of nonverbal behavior: Categories, origins, usage, and coding', *Semiotica* **1**(1), 49–98.

Electronic Design (1966), number v. 14, nos. 1, Hayden Publishing Company.

Engelbart, D. C. & English, W. K. (1968), A research center for augmenting human intellect, *in* 'Proceedings of the December 9-11, 1968, fall joint computer conference, part I', ACM, pp. 395–410.

Esparza, G. (2007), 'Urban parasites', `http://www.parasitosurbanos.com/`. Accessed: 2016-03-17.

Evans, J., Krishnamurthy, B., Pong, W., Croston, R., Weiman, C. & Engelberger, G. (1989), 'Helpmate: A robotic materials transport system', *Robotics and Autonomous Systems* **5**(3), 251–256.

Feil-Seifer, D. & Matarić, M. J. (2011), 'Socially assistive robotics', *Robotics & Automation Magazine, IEEE* **18**(1), 24–31.

Festo, A. & Co, K. (2009), 'Bionic learning network', *`http: // www. festo. com/ cms/ de_ de/ 4981. htm`* .

Fischer, M. (2009*a*), 'Airpenguin', *Festo AG* .

Fischer, M. (2009*b*), 'Aquapenguin, a biomechatronic overall concept', *Festo AG* .

Fukui, K., Nishikawa, K., Ikeo, S., Honda, M. & Takanishi, A. (2006), Development of a human-like sensory feedback mechanism for an anthropomorphic talking robot, *in* 'Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on', IEEE, pp. 101–106.

Graf, B., Hans, M. & Schraft, R. D. (2004), 'Care-o-bot iidevelopment of a next generation robotic home assistant', *Autonomous robots* **16**(2), 193–205.

Greer, J. A. (2014), Building emotional authenticity between humans and robots, *in* 'Workshops in ICSR 2014'.

Grennan, K. (2011), 'Prototype robot armpit', `http://www.kevingrennan.com/prototype-robot-armpit/`. Accessed: 2016-03-17.

Grimshaw, M. (2009), ‘The audio uncanny valley: Sound, fear and the horror game’.

Hailstone, J. C., Omar, R., Henley, S. M., Frost, C., Kenward, M. G. & Warren, J. D. (2009), ‘It’s not what you play, it’s how you play it: Timbre affects perception of emotion in music’, *The quarterly Journal of Experimental psychology* **62**(11), 2141–2155.

Hall, E. T. (1966), *The hidden dimension*, Doubleday & Co.

Hansmeyer, M. et al. (2010), Design by subdivision, *in* ‘Proceedings of Bridges 2010: Mathematics, Music, Art, Architecture, Culture’, Tessellations Publishing, pp. 167–174.

Hermann, T., Drees, J. M. & Ritter, H. (2003), ‘Broadcasting auditory weather reports-a pilot project’, *Georgia Institute of Technology* .

Hertz, G. (2008), ‘Cockroach controlled mobile robot’, `http://www.conceptlab.com/roachbot/`. Accessed: 2016-03-17.

Hine, B., Stoker, C., Sims, M., Rasmussen, D., Hontalas, P., Fong, T., Steele, J., Barch, D., Andersen, D., Miles, E. et al. (1994), The application of telepresence and virtual reality to subsea exploration, *in* ‘Second Workshop on Mobile Robots for Subsea Environments’.

Hoggett, R. (2010), ‘Cybernetic zoo’, `http://cyberneticzoo.com/?p=2146`. Accessed: 2016-06-24.

Horowitz, S. S. (2012*a*), ‘The science and art of listening’, *New York Times* **9**.

Horowitz, S. S. (2012*b*), *The universal sense: How hearing shapes the mind*, Bloomsbury Publishing USA.

Huron, D. (2008), ‘A comparison of average pitch height and interval size in major- and minor-key themes: Evidence consistent with affect-related pitch prosody’.

Huron, D., Anderson, N. & Shanahan, D. (2014), ‘you cant play a sad song on the banjo: acoustic factors in the judgment of instrument capacity to convey sadness’, *Empirical Musicology Review* **9**(1), 29–41.

Huron, D., Kinney, D. & Precoda, K. (2006), ‘Influence of pitch height on the perception of submissiveness and threat in musical passages’.

Ip, G. W.-m., Chiu, C.-y. & Wan, C. (2006), 'Birds of a feather and birds flocking together: physical versus behavioral cues may lead to trait-versus goal-based group perception.', *Journal of personality and social psychology* **90**(3), 368.

Jansen, T. (2008), 'Strandbeests', *Architectural Design* **78**(4), 22–27.

Jee, E.-S., Jeong, Y.-J., Kim, C. H. & Kobayashi, H. (2010), 'Sound design for emotion and intention expression of socially interactive robots', *Intelligent Service Robotics* **3**(3), 199–206.

Jee, E.-S., Kim, C. H., Park, S.-Y. & Lee, K.-W. (2007), Composition of musical sound expressing an emotion of robot based on musical factors, *in* 'Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on', IEEE, pp. 637–641.

Jee, E.-S., Park, S.-Y., Kim, C. H. & Kobayashi, H. (2009), Composition of musical sound to express robot's emotion with intensity and synchronized expression with robot's behavior, *in* 'Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on', IEEE, pp. 369–374.

Jibo (2015), 'Jibo - the worlds first social robot', `https://www.jibo.com/?utm_campaign=brand&utm_source=adwords&utm_medium=introJibo-a&gclid=CNXy4Z7EwcOCFYqPfgodvpUMNA`. Accessed: 2016-06-24.

Johnstone, T., Van Reekum, C. M. & Scherer, K. R. (2001), 'Vocal expression correlates of appraisal processes', *Appraisal processes in emotion: Theory, methods, research* pp. 271–284.

Kaufman, S. B. & Van Ellin, M. (1995), 'Hanc: A case study in invention', *Ageing International* **22**(1), 26–28.

Kempelen, W., Brekle, H. E. & Wildgen, W. (1970), *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*, Friedrich Frommann Verlag.

Khan, Z. (1998), 'Attitudes towards intelligent service robots', *NADA KTH, Stockholm* **17**.

Kidd, C. & Breazeal, C. (2005), 'Comparison of social presence in robots and animated characters', *Interaction Journal Studies* .

Kiesler, S., Powers, A., Fussell, S. R. & Torrey, C. (2008), 'Anthropomorphic interactions with a robot and robot-like agent', *Social Cognition* **26**(2), 169–181.

Köhler, W. (1929), *Gestalt psychology*, Liveright.

Komatsu, T. (2005), Toward making humans empathize with artificial agents by means of subtle expressions, *in* 'Affective Computing and Intelligent Interaction', Springer, pp. 458–465.

Komatsu, T. & Yamada, S. (2011), 'How does the agents' appearance affect users' interpretation of the agents' attitudes: Experimental investigation on expressing the same artificial sounds from agents with different appearances', *Intl. Journal of Human–Computer Interaction* **27**(3), 260–279.

Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K. & Nakano, M. (2010), Artificial subtle expressions: intuitive notification methodology of artifacts, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM, pp. 1941–1944.

Kozima, H., Michalowski, M. P. & Nakagawa, C. (2009), 'Keepon', *International Journal of Social Robotics* **1**(1), 3–18.

Kristoffersson, A., Coradeschi, S. & Loutfi, A. (2013), 'A review of mobile robotic telepresence', *Advances in Human-Computer Interaction* **2013**, 3.

Larsson, P. (2010), Tools for designing emotional auditory driver-vehicle interfaces, *in* 'Auditory Display', Springer, pp. 1–11.

Le Groux, S. & Verschure, P. (2010), Emotional responses to the perceptual dimensions of timbre: A pilot study using physically informed sound synthesis, *in* 'Proc. 7th Int. Symp. Comput. Music Model', pp. 1–15.

Lee, K. M., Jung, Y., Kim, J. & Kim, S. R. (2006), 'Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction', *International Journal of Human-Computer Studies* **64**(10), 962–973.

Léger, F. & Murphy, D. (1924), *Ballet mécanique*, Synchro-cine.

Lehni, J. (2008), 'Soft monsters', *Perspecta* **40**, 22–27.

Lehrman, P. D. & Singer, E. (2008), Doing good by the bad boy: Performing george antheils ballet mécanique with robots, *in* 'Technologies for Practical Robot Applications, 2008. TePRA 2008. IEEE International Conference on', IEEE, pp. 13–18.

Leichtentritt, H. (1934), 'Mechanical music in olden times', *The Musical Quarterly* **20**(1), 15–26.

Levy, S. (2001), *Hackers: Heroes of the computer revolution*, Vol. 4, Penguin Books New York.

Leyzberg, D., Avrunin, E., Liu, J. & Scassellati, B. (2011), Robots that express emotion elicit better human teaching, *in* 'Proceedings of the 6th international conference on Human-robot interaction', ACM, pp. 347–354.

Li, J. (2015), 'The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents', *International Journal of Human-Computer Studies* **77**, 23–37.

Li, X., MacDonald, B. & Watson, C. I. (2009), Expressive facial speech synthesis on a robotic platform, *in* 'Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on', IEEE, pp. 5009–5014.

Lieberman, P. & Michaels, S. B. (1962), 'Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech', *The Journal of the Acoustical Society of America* **34**(7), 922–927.

Lövheim, H. (2012), 'A new three-dimensional model for emotions and monoamine neurotransmitters', *Medical hypotheses* **78**(2), 341–348.

Mori, M., MacDorman, K. F. & Kageki, N. (2012), 'The uncanny valley [from the field]', *Robotics & Automation Magazine, IEEE* **19**(2), 98–100.

Morris, J. D. (1995), 'Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response', *Journal of advertising research* **35**(6), 63–68.

Murray, I. R. & Arnott, J. L. (1993), 'Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion', *The Journal of the Acoustical Society of America* **93**(2), 1097–1108.

Niculescu, A., van Dijk, B., Nijholt, A., Li, H. & See, S. L. (2013), 'Making social robots more attractive: the effects of voice pitch, humor and empathy', *International journal of social robotics* **5**(2), 171–191.

Nourbakhsh, I. R., Bobenage, J., Grange, S., Lutz, R., Meyer, R. & Soto, A. (1999), 'An affective mobile robot educator with a full-time job', *Artificial Intelligence* **114**(1), 95–124.

Ohala, J. (1996), 'The frequency code underlies the sound symbolic use of voice pitch', *Sound Symbolism* pp. 325–347.

Ohala, J. J. (1980), 'The acoustic origin of the smile', *The Journal of the Acoustical Society of America* **68**(S1), S33–S33.

Ohala, J. J. (1983), 'Cross-language use of pitch: an ethological view', *Phonetica* **40**(1), 1–18.

Ohala, J. J., Hinton, L. & Nichols, J. (1997), Sound symbolism, *in* 'Proc. 4th Seoul International Conference on Linguistics [SICOL]', pp. 98–103.

Opfer, J. E. & Gelman, S. A. (2010), 'Development of the animate-inanimate distinction', *The Wiley-Blackwell Handbook of Childhood Cognitive Development,* pp. 213–238.

Orellana, F. (1999), 'the hive', `http://fernandoorellana.com/projects/the-hive/`. Accessed: 2016-03-17.

Orellana, F. (2007), 'Elevators music', `http://fernandoorellana.com/projects/elevators-music/`. Accessed: 2016-03-17.

Oviatt, S. (1997), 'Multimodal interactive maps: Designing for human performance', *Human-computer interaction* **12**(1), 93–129.

Oviatt, S., Coulston, R. & Lunsford, R. (2004), When do we interact multimodally?: cognitive load and multimodal communication patterns, *in* 'Proceedings of the 6th international conference on Multimodal interfaces', ACM, pp. 129–136.

Pangburn, D. (2016), 'This dapper robot is an art critic', `http://thecreatorsproject.vice.com/blog/robot-art-critic-berenson`. Accessed: 2016-03-17.

Pauline, M. (1979), 'Survival research laboratories', `http://www.srl.org/`. Accessed: 2016-03-17.

Penny, S. (2011), 'Simon penny', `http://simonpenny.net/`. Accessed: 2016-03-17.

Pierre-Yves, O. (2003), 'The production and recognition of emotions in speech: features and algorithms', *International Journal of Human-Computer Studies* **59**(1), 157–183.

Pieskä, S., Luimula, M., Jauhiainen, J. & Spiz, V. (2012), 'Social service robots in public and private environments', *Recent Researches in Circuits, Systems, Multimedia and Automatic Control* pp. 190–196.

Pollack, M. E., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., Engberg, S., Matthews, J. T., Dunbar-Jacob, J., McCarthy, C. E. et al. (2002), Pearl: A mobile robotic assistant for the elderly, *in* 'AAAI workshop on automation as eldercare', Vol. 2002, pp. 85–91.

Post, O. & Huron, D. (2009), 'Western classical music in the minor mode is slower (except in the romantic period)'.

Prade, E. L. (2002), 'The early days of eat', *MultiMedia, IEEE* **9**(2), 4–5.

Raibert, M., Blankespoor, K., Nelson, G., Playter, R. & Team, T. (2008), Bigdog, the rough-terrain quadruped robot, *in* 'Proceedings of the 17th World Congress', Vol. 17, pp. 10822–10825.

Rane, P., Mhatre, V. & Kurup, L. (2014), Study of a home robot: Jibo, *in* 'International Journal of Engineering Research and Technology', Vol. 3, ESRSA Publications.

Ray, C., Mondada, F. & Siegwart, R. (2008), What do people expect from robots?, *in* 'Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on', IEEE, pp. 3816–3821.

Read, R. (2014), *A study of non-linguistic utterances for social human-robot interaction*, Plymouth University.

Read, R. & Belpaeme, T. (2012), How to use non-linguistic utterances to convey emotion in child-robot interaction, *in* 'Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction', ACM, pp. 219–220.

Read, R. & Belpaeme, T. (2013), People interpret robotic non-linguistic utterances categorically, *in* 'Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction', IEEE Press, pp. 209–210.

Reeves, B. & Nass, C. (1996), *How people treat computers, television, and new media like real people and places*, CSLI Publications and Cambridge university press.

Reichardt, J. (1969), *Cybernetic serendipity: the computer and the arts*, Praeger.

Rinaldo, K. (2015), 'Ken rinaldo', `http://www.kenrinaldo.com/`. Accessed: 2016-03-17.

Roehling, S., MacDonald, B. & Watson, C. (2006), Towards expressive speech synthesis in english on a robotic platform, *in* 'Proceedings of the Australasian International Conference on Speech Science and Technology', pp. 130–135.

Royakkers, L. & van Est, R. (2015), 'A literature review on new robotics: automation from love to war', *International journal of social robotics* **7**(5), 549–570.

Russell, J. A. (1980), 'A circumplex model of affect.', *Journal of personality and social psychology* **39**(6), 1161.

Russell, J. A. & Mehrabian, A. (1977), 'Evidence for a three-factor theory of emotions', *Journal of research in Personality* **11**(3), 273–294.

Saerbeck, M. & Bartneck, C. (2010), Perception of affect elicited by robot motion, *in* 'Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction', IEEE Press, pp. 53–60.

Sakagami, Y., Watanabe, R., Aoyama, C., Matsunaga, S., Higaki, N. & Fujimura, K. (2002), The intelligent asimo: System overview and integration, *in* 'Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on', Vol. 3, IEEE, pp. 2478–2483.

Sapaty, P. (2015), 'Military robotics: Latest trends and spatial grasp solutions', *International Journal of Advanced Research in Artificial Intelligence* **4**(4).

Sawada, H. (2007), *Talking robot and the autonomous acquisition of vocalization and singing skill*, INTECH Open Access Publisher.

Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S. & Tow, R. (2002), Experiences with sparky, a social robot, *in* 'Socially Intelligent Agents', Springer, pp. 173–180.

Scherer, K. R., Johnstone, T. & Klasmeyer, G. (2003), 'Vocal expression of emotion', *Handbook of affective sciences* pp. 433–456.

Scherer, K. R., London, H. & Wolf, J. J. (1973), 'The voice of confidence: Paralinguistic cues and audience evaluation', *Journal of Research in Personality* **7**(1), 31–44.

Scheutz, M., Schermerhorn, P., Kramer, J. & Anderson, D. (2007), 'First steps toward natural human-like hri', *Autonomous Robots* **22**(4), 411–423.

Schutz, M., Huron, D., Keeton, K. & Loewer, G. (2008), 'The happy xylophone: acoustics affordances restrict an emotional palate'.

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C. & Young, J. E. (2015), Poor thing! would you feel sorry for a simulated robot?: A comparison of empathy toward a physical and a simulated robot, *in* 'Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction', ACM, pp. 125–132.

Shi, C., Shiomi, M., Kanda, T., Ishiguro, H. & Hagita, N. (2015), 'Measuring communication participation to initiate conversation in human–robot interaction', *International Journal of Social Robotics* pp. 1–22.

Siegman, A. W. & Boyle, S. (1993), 'Voices of fear and anxiety and sadness and depression: the effects of speech rate and loudness on fear and anxiety and sadness and depression.', *Journal of Abnormal Psychology* **102**(3), 430.

Skinner, B. F. (1951), *How to teach animals*, Freeman.

Smith, J. O. (2010), *Physical Audio Signal Processing*, `http://ccrma.stanford.edu/~jos/pasp/`.

Snow, M. (1972), 'La région centrale', *Cinema Canada* .

Sobin, C. & Alpert, M. (1999), 'Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy', *Journal of psycholinguistic research* **28**(4), 347–365.

Streeter, L. A., Macdonald, N. H., Apple, W., Krauss, R. M. & Galotti, K. M. (1983), 'Acoustic and perceptual indicators of emotional stress', *Journal of the Acoustical Society of America* **73**(4), 1354–t360.

Sutherland, I. E. (1960), 'Stability in steering control', *Electrical Engineering* **79**(4), 298–301.

Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R. & Hayashi, K. (2015), Pepper learns together with children: Development of an educational application, *in* 'Humanoid Robots (Humanoids), 2015 IEEE-RAS 15th International Conference on', IEEE, pp. 270–275.

Tartter, V. C. (1980), 'Happy talk: Perceptual and acoustic effects of smiling on speech', *Perception & psychophysics* **27**(1), 24–27.

Tresset, P. & Leymarie, F. F. (2013), 'Portrait drawing by paul the robot', *Computers & Graphics* **37**(5), 348–363.

Tsui, K. M., Desai, M., Yanco, H. A. & Uhlik, C. (2011), Exploring use cases for telepresence robots, *in* 'Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on', IEEE, pp. 11–18.

Turner, B. & Huron, D. (2008), 'A comparison of dynamics in major-and minor-key works'.

Völker, N. (2009), 'Makers & spectators', `http://www.nilsvoelker.com/content/mu/`. Accessed: 2016-03-17.

Wadlow, T. A. (1981), 'The xerox alto computer', *Byte Magazine* **6**(9), 58–68.

Wainer, J., Feil-Seifer, D. J., Shell, D., Matarić, M. J. et al. (2006), The role of physical embodiment in human-robot interaction, *in* 'Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on', IEEE, pp. 117–122.

Walter, W. G. (1950), 'An imitation of life'.

Wang, Y., Laby, K. P., Jordan, C. S., Butner, S. E. & Southard, J. (2005), 'Medical tele-robotic system'. US Patent 6,925,357.

Williams, C. E. & Stevens, K. N. (1972), 'Emotions and speech: Some acoustical correlates', *The Journal of the Acoustical Society of America* **52**(4B), 1238–1250.

Wu, Y.-H., Fassert, C. & Rigaud, A.-S. (2012), 'Designing robots for the elderly: appearance issue and beyond', *Archives of gerontology and geriatrics* **54**(1), 121–126.

Yeo, S. (2015), 'Robots collaboration', `http://www.shihyunyeo.com/index.php/13-works/53-robots-collaboration`. Accessed: 2016-03-17.

Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E. & Lefeber, D. (2011), Emogib: emotional gibberish speech database for affective human-robot interaction, *in* 'Affective Computing and Intelligent Interaction', Springer, pp. 163–172.

Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E. & Lefeber, D. (2013), Multi-modal emotion expression for affective human-robot interaction, *in* 'Proceedings of the Workshop on Affective Social Speech Signals (WASSS 2013), Grenoble, France'.

Yilmazyildiz, S., Latacz, L., Mattheyses, W. & Verhelst, W. (2010), Expressive gibberish speech synthesis for affective human-computer interaction, *in* 'Text, Speech and Dialogue', Springer, pp. 584–590.

Yilmazyildiz, S., Mattheyses, W., Patsis, Y. & Verhelst, W. (2006), Expressive speech recognition and synthesis as enabling technologies for affective robot-child communication, *in* 'Advances in Multimedia Information Processing-PCM 2006', Springer, pp. 1–8.

Zivanovic, A. (2007), 'The senster', `http://www.senster.com/ihnatowicz/index.htm`. Accessed: 2016-06-24.