MACHINE LEARNING DRIVEN EMOTIONAL MUSICAL PROSODY FOR HUMAN-ROBOT INTERACTION

A Dissertation
Presented to
The Academic Faculty

By

Richard Savery

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Music

Georgia Institute of Technology

December 2021

MACHINE LEARNING DRIVEN EMOTIONAL MUSICAL PROSODY FOR HUMAN-ROBOT INTERACTION

Approved by:

Dr. Gil Weinberg, Advisor
School of Music
Georgia Institute of Technology

Dr. Claire Arthur
School of Music
Georgia Institute of Technology

Dr. Jason Free
School of Music
Georgia Institute of Technology

Dr. Ayanna Howard
College of Engineering
Ohio State University

Date Approved: October 28, 2021

ACKNOWLEDGEMENTS

I would like to thank my advisor Gil Weinberg for all his support and direction throughout my PhD. Gil's guidance has truly helped develop my research and changed my approach to music technology. Looking back across my PhD I feel like my work has improved significantly, largely through Gil's influence. I feel very fortunate to have had the chance to work with Gil over the last few years and look forward to continuing into the future.

I am grateful to my committee Claire Arthur, Jason Freeman, and Ayanna Howard all of whom have helped shape and drastically improve the contents of this dissertation and draw everything together. Jason has also been a constant supporter from the start of my Phd, and I am sincerely thankful for his role in graduate studies.

This dissertation has been supported by the National Science Foundation, National Robotics Initiative (Grant 1925178). This support has allowed the research to greatly develop beyond what would have otherwise been possible, allowing exploration with a wider range of robots and musicians.

I would also like to thank three of the collaborators on the research in this dissertation, Ryan Rose, Lisa Zahray and Amit Rogel all of whom have been central to the overall research. Ryan was key to the initial development of the concept of EMP, and created the gestures and hardware implementation for Shimi. Lisa played a key role in initial studies and dataset validation as well as improving the conceptual ideas of EMP. Most recently Amit has developed the gestures for the robotic arms, designing incredible arm interaction for human collaborators, while improving the experimental design.

I have been very privileged to have worked with so many amazing collaborators in the Robotic Musicianship Lab and GTCMT especially Ning Yang, Raghavasimhan Sankaranarayanan, Mike Winters, Zachary Kondak, Mohammad Jafari, Yilin Zhang, Madhukesh Ayyagari, Jason Smith and Benjie Genchel as well as many others. I've been very fortunate

to be part of a thriving community at GTCMT which has helped expand my own directions and acted as a continual influence. I've also had the chance to work with many extremely talented undergraduates through the VIP program, whose work has contributed to parts of this dissertation, and who have motivated me to continually improve my own research.

I would like to thank my parents and brother who have made countless trips to the US over the last 8 years, as well as my grandparents and extended family who have always embraced my endeavours. To my wife Anna who has been an endless supporter, moving with me to Atlanta, reading papers late at night and always ready to listen to a plan for future study. I aspire to be as supportive during her own PhD next year. Finally, to my daughters Keira and Zoe who have always supported my trips to see the robots and who's own creativity and thought process have been a constant inspiration to all my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation presents a method for non-anthropomorphic human-robot interaction using a newly developed concept entitled Emotional Musical Prosody (EMP). EMP consists of short expressive musical phrases capable of conveying emotions, which can be embedded in robots to accompany mechanical gestures. The main objective of EMP is to improve human engagement with, and trust in robots while avoiding the uncanny valley. We contend that music - one of the most emotionally meaningful human experiences - can serve as an effective medium to support human-robot engagement and trust. EMP allows for the development of personable, emotion-driven agents, capable of giving subtle cues to collaborators while presenting a sense of autonomy.

We present four research areas aimed at developing and understanding the potential role of EMP in human-robot interaction. The first research area focuses on collecting and labeling a new EMP dataset from vocalists, and using this dataset to generate prosodic emotional phrases through deep learning methods. Through extensive listening tests, the collected dataset and generated phrases were validated with a high level of accuracy by a large subject pool. The second research effort focuses on understanding the effect of EMP in human-robot interaction with industrial and humanoid robots. Here, significant results were found for improved trust, perceived intelligence, and likeability of EMP enabled robotic arms, but not for humanoid robots. We also found significant results for improved trust in a social robot, as well as perceived intelligence, creativity and likeability in a robotic musician.

The third and fourth research areas shift to broader use cases and potential methods to use EMP in HRI. The third research area explores the effect of robotic EMP on different personality types focusing on extraversion and neuroticism. For robots, personality traits offer a unique way to implement custom responses, individualized to human collaborators. We discovered that humans prefer robots with emotional responses based on high extraversion and low neuroticism, with some correlation between the humans collaborator's own personality traits. The fourth and final research question focused on scaling up EMP to support interaction between groups of robots and humans. Here, we found that improvements in trust and likeability carried across from single robots to groups of industrial arms. Overall, the thesis suggests EMP is useful for improving trust and likeability for industrial, social and robot musicians but not in humanoid robots. The thesis bears future implications for HRI designers, showing the extensive potential of careful audio design, and the wide range of outcomes audio can have on HRI.

CHAPTER 1
INTRODUCTION

Many recent robotic systems focus on achieving human-like features and interactions through anthropomorphic design. There is a growing body of work that suggests the counter approach of mechanomorphic design, has a greater potential for improved interaction, through higher levels of trust and engagement with human collaborators [1]. The concept of "uncanny valley" describes that as a robot becomes more human, they become more appealing, until they reach a point where they elicit revulsion. To address this negative effect, Moore suggests the contrasting, mechanomorphic, "canny" approach, whereby robots are developed that are clearly robots [2] and openly display their capabilities and robotic features. We propose that using emotional musical prosody (EMP) offers a "canny" approach to robot design and communication for social and industrial applications, leveraging mechanomorphic design for improved interaction. We hypothesize that EMP can improve human-robot trust, likeability, perceived intelligence, and task performance on multiple robotic platforms. This dissertation first describes a new model for creating EMP, and then presents studies to understand the potential application of EMP across robotic platforms.

1.1 Motivation

As co-robots become more prevalent at home, work, and public environments, a need arises for improved modes of communication between humans and groups of robots. A meta-study of human-robot trust [3] has shown that robot-related attributes are the main contributors to building trust in Human-Robot-Interaction, affecting trust more than environmental and human related factors. One of the key robotic attributes shown to contribute toward trust building with humans is "Robot Personality". Related research on artificial agents and personality traits [4, 5] indicates that an effective approach for building trust and other collaboration metrics with artificial agents is through conveying emotions using subtle non-verbal communication channels such as prosody and gesture. These channels can help convey intentions as well as expressions including humor, sarcasm, irony, and state-of-mind, which help build social relationship and trust.

While synthesized linguistic speech has seen great advances in recent years [6], it has been shown that humans can be attentive to only limited numbers of linguistic channels, which can lead to the loss of important information. Facial expression, an alternate non-semantic emotion-carrying modality [7], bears the risk of creating eeriness and revulsion known and often involves a large number of Degrees of Freedom (DoFs) that cannot easily scale to support large groups of robots [8].

In speech and language literature, prosody is clearly defined as the features of speech that are non-semantic, including the intonation, tone, stress, rhythm and pitch [9]. These features of speech have been extended to describe prosody as "the music of everyday speech" [10]. *Musical prosody* has a wider range of interpretations; Palmer and Hutchins describe that musical prosody is expressive, emotional musical phrases, inspired by prosodic speech features [11]. They emphasize that musical prosody is related to musical expression, "because, as in speech, performers manipulate music for certain expressive and coordinating functions". For the purpose of this dissertation, we build on the definition proposed by Palmer and Hutchins to label EMP as musical phrases designed to portray an emotion through variation in pitch, rhythm, timbre, intonation and stress.

Recent efforts to generate and manipulate prosody focused on linguistic robotic communication [12], and have been successful in conveying expressions such as approval, prohibition, attention, and comfort [13]. However, to our knowledge, there has been no successful effort to use EMP without language for robot interaction. We believe that EMP without language can support many interactions where language is not needed, such as in noisy environments or group collaborations, allowing humans to execute other tasks while receiving background information. To address this goal, we propose the use of algorithmic music analysis and generation to drive a novel EMP generator. Music is a powerful medium to convey emotions [14], and shares many of the underlying building blocks of prosody such as pitch, timing, loudness, intonation, and timbre [10, 15, 16]. We, therefore, propose that using musical prosody as a means of communication has a transformative potential for the field of human-robot interaction.

1.2 Research Questions

In order to explore the effect EMP can have on HRI, our central goal was to analyze the broad question:

> **What effects can a generative system for EMP have on trust, likeability and perceived intelligence in individual and group human-robot interaction?**

We expanded this central question into four research questions to address multiple robotic platforms, users and use cases. To realize the potential of EMP, we first collected and a validated a new Emotionally-Labeled-Musical-Prosody (ELMP) dataset. We then used the dataset to generate EMP that is able to convey expressive non-semantic information to humans. This was followed by multiple studies considering the use of EMP in many different implementations and formats. We aimed primarily to understand how EMP could improve metrics such as trust and likeability, and how these vary between different platforms. EMP was then embedded in a personality driven robot, using different emotional responses to stimuli, to imply varying robotic personalities. We concluded by studying the effect of EMP in a group of three robots with one human participant.

1.2.1 EMP Generation and Dataset Collection

RQ 1: Can a data driven, EMP system generate musical phrases that can be labelled by listeners?

To develop a EMP generator we first collected the ELMP dataset, consisting of audio from three vocalists. Each vocalist improvised for four hours, broken up into 15 minute segments for 20 discrete emotion classes. The 20 discrete classes were taken from the Geneva Emotion Wheel, which uses 5 categories from each quadrant of the circumplex model [17]. We then conducted multiple classification tasks to clarify machine learning models ability to separate the recorded emotions using audio features. This was followed by an extensive listening test with users, to identify how well the emotions identified by each vocalist could be tagged by a broader audience.

We created a model that could generate EMP in real-time. This system was designed foremost to allow rapid generation and dialogue exchange between humans and robots. For this reason the system combined symbolic deep learning using MIDI and annotated timbre features, through a Conditional Convolution Variational Auto-encoder, with an emotion-tagged audio sampler using generated feature descriptors. We evaluated this system primarily through listening tests with participants choosing the emotion they believe the system is attempting to display. We tested the hypothesis that the generative musical system would result in users correctly labeling emotions with an similar accuracy rate as that of the (human-generated) ELMP dataset.

1.2.2 EMP, Trust and HRI Metrics

RQ 2: How does EMP alter the level of likeability, perceived intelligence and trust in social, industrial, humanoid and robotic musicians?

The second research question focuses on using the generated EMP in multiple robotic platforms. We analyzed EMP implementations in a robotic industrial arm, a humanoid model, the social robot Shimi, and the musical robot Shimon. For the industrial arm we conducted a study analyzing potential benefits of using EMP to allow the robot to respond emotionally to a human's actions. We tested participants' responses to interacting with a virtual robot arm that acted as a decision agent, helping participants select the next number in a sequence. We then compared results from three versions of the application in a between-group experiment, where the robot presented different emotional reactions to the user's input depending on whether the user agreed with the robot and whether the user's choice was correct. One version used EMP audio phrases selected from our dataset of singer improvisations, the second version used audio consisting of a single pitch randomly assigned to each emotion, and the final version used no audio, only gestures. In each version, the robot reacted with emotional gestures. Participants completed a trust survey following the interaction, which we then compared to how often they followed suggestions in the simulation. We then replicated this study for the humanoid robot to analyse differences across platforms.

The next study used the social robot Shimi interacting with emotional phrases that match a users' choice emotion in an in-person study. This study was conducted as a between-group study with one version of Shimi using EMP while the other used a SOTA text-to-speech system and trust measured in a post-interaction 40 question survey. The final study analyzed EMP in the musical robot Shimon. We conducted a between-groups study with participants watching a musician interact for 30 seconds with Shimon, followed by Shimon responding either with EMP or text-to-speech. We collected survey responses for likeability, perceived intelligence, animacy and anthropomorphism. We also aimed to understand if EMP is capable of altering the perception of a robot's key functionality, in this case ratings of musical generation ability and creativity.

1.2.3 Personality Preferences

RQ 3: Does a person's personality alter their ratings of different emotional responses portrayed through robotic EMP?

Research question 1 focused on developing a EMP generation system, while research question 2 analyzed EMP's impact on multiple robotic systems. Research question 3 further explores the potential of EMP when combined with personality traits to understand how emotional reactions can be leveraged in human-robot interaction. Personality has been widely utilized in human robotic interaction research, such as in work that embed human personality in a robot to drive certain reactions and uses [18]. While emotion is considered a critical feature of personality and is intertwined with the definition of personality itself [19], little research has been conducted addressing the interaction of personality, emotion, and robotics. For this research question we considered links between two of the Big Five personality types, Neuroticism and Extraversion. We chose Neuroticism and Extraversion as both have shown robust and consistent findings in their role in emotion for a human's personality [20].

To display these personality traits, we programmed robots to show different emotional responses to stimuli through EMP and gesture, based on common human emotion responses. These emotional responses were reduced to the robotic arm displaying responses to visual stimuli, for example High Neuroticism and low Extraversion (HighN-LowE) personalities are consistently more likely to respond to positive stimuli with lower valence emotions, such as relief. Low Neuroticism and High Extraversion (LowN-HighE) personalities are much more likely to respond directly with Joy or Happiness [21].

We conducted two separate studies aiming to evaluate our EMP for future use of emotion variation and personality. The first study compared preferences between a robot and human personality types. We used a between-group study, with either a high Neuroticism and low Extraversion robot or low Neuroticism and high Extraversion robot. Robot personality traits were displayed through varying emotional responses demonstrated in to visual stimuli. We then compared preferences from different human personality types for each robot version. The second study analyzed whether the same preference would occur when choosing between isolated emotional responses, using a within-group study design.

1.2.4 Group Interaction

RQ 4: Can EMP be scaled to group robotics, to reduce entitativity while increasing trust and likeability ratings?

Research questions 1, 2, and 3 addressed dyadic human-robot interaction with a single human and single robot participant. Emotions, however, are an inherently social activity that are elicited by others, expressed towards groups, and regulated to influence people [22]. Research in human-robot interaction has focused on the relationship between a single robot and a single human participant. Only limited research has addressed the contrasting dynamic when humans interact with a group of robots. This dynamic adds additional human-robot interaction considerations, such as the level of entitativity, which is the identification of a group as a single entity as opposed to a collection of individuals. Research Question 4 examined the role EMP can play in improving the interaction between humans and groups of robots by modifying the level of entitativity.

We conducted a between-group experiment, presenting to subjects a group of industrial robotic arms performing a task either without sound, with the same EMP voice for each robot, or with contrasting voices for different robots. We then analyzed the same metrics used in individual robotics, such as trust and likeability and measured if the improvements from individual robots carries to groups of robots. This was follow be an analysis to determine if through subtle variations in timbre, EMP is able to alter the level of entitativity perceived by external observers. Finally, we aimed to extend broader HRI understandings of the interaction between entitativity and common HRI metrics.

1.3 Contributions

This dissertation will lead to new knowledge about the analysis and generation of non-semantic emotional communication channels and their effect on human-robot interaction. The project will, for the first time, show how a novel musically-informed generative model that manipulates emotional content in robotic communication could be integrated into other emotion conveying channels such as physical gestures to improve and enrich human-robot interaction. The project will lead to new knowledge about the relationship between musical features and vocal prosody. It will provide new insights regarding the potential and limitations of non-semantic emotional communication channels constrained by limited degrees of freedom (DoFs), velocity restrictions, and non-humanoid design. By manipulating robotic emotional conveyance in a well controlled environment, the project will also lead to novel insights on human emotional response to emotional robots. The project will therefore provide a new paradigm for increasing engagement, relatability, and trust in various scales of human-robot interactions. The primary contributions are listed below:

1. A new model for EMP generation for robotics. Informed by machine learning, this model will apply musical features over vocal prosody [Chapter 3]

2. A new dataset of emotion-tagged musical phrases [Chapter 3]

3. A novel approach for building trust between humans and robots through different combinations of non-semantic emotional conveyance [Chapter 4]

4. New knowledge about humans' preference for robotic emotional response based on personality [Chapter 5]

5. New knowledge on applications for EMP in groups of robots [Chapter 6]

CHAPTER 2

RELATED WORK

2.1 Human-Robot Interaction

2.1.1 Methods of Communication

Verbal language based interaction is the prominent form of audio communication used in human-robot interaction [23] covering a wide range of tasks from robot companions [24] to industry [25]. With the growth of text-to-speech platforms and voice assistants, verbal communication is now easy to implement across many platforms [26]. It has been shown that the voice quality and characteristics affect the nature of the interaction, with playful voices changing the tone of the dialogue [27]. Many voice systems aim to sound as human-like as possible, however studies have shown that effort to sound like a human can negatively affect the expectations set by the end user and can decrease usage [28]. For example, it has been suggested that the dominant form of a female voice widely accepted by Amazon Alexa, Apple's Siri and Google Voice is not necessarily the best form of communication especially in robotics [29].

Often, human-robot interactions do not include language. Jones divided these non-verbal forms of communication into six categories, kinesics, proxemics, haptics, chronemics, vocalics, and presentation [30, 31]. Kinesics includes communication through body movement, such as gestures [32], or facial expressions, while proxemics focuses on the robotic positioning in space, such as the distance from a human collaborator [33]. Haptics refers to touch based methods [34], while chronemics includes subtle traits such as hesitation [35]. Presentation includes the way the robot appears, such as changes based on different behaviour [36]. The category, vocalics, includes methods such as prosody [12]. The vast majority of these communication techniques require significant technical and financial expense and variation to a system [31].

Social robot communication modalities are often derived from human behaviour such as gestures and gaze [37]. However, these modalities are not readily available to robotic arms [38]. Amongst research into methods for robotic arms to communicate and signal their intent, there is no standardized set of approaches [39]. While controlling movement to show intent has shown successful results [40], changes to path planning and movement dynamics is often not feasible without altering the core intent of carefully planned movement. Another effective method for arms to display their intent is through artificial vision of a robot's future trajectory, such as a human worn augmented reality display [41]. However, this requires a significant financial investment and is a potential distraction to the user.

There is only limited work in sound and HRI outside the use of speech systems, with research on the impact of sound relatively rare [42]. Studies have been conducted to analyze whether the use of a beep improves perception of a robot with positive results, although more considered application of the range of possible audio sounds has not been conducted [43]. Consequential sounds are the sounds made by a robot in normal operation, such as motors and physical movement. The sound from motors has been used as a communication tool through modification of gesture [44], as well as used to improve localization [45]. Overall, consequential sounds have been analyzed for their impact on interaction with primarily negative results [46, 47]. Sonification of robotic movements has been examined, such as in relation to emotions for children with Autism Spectrum Disorder (ASD) [48], or for general movement of robots [49]. While there are multiple attempts to incorporate sound beyond spoken language into robotics, it is ultimately very limited in scope with broad potential for further research. There has not been the same sound tested on multiple platforms, or even the same platform in different interaction types, and each sound implementation is very rarely explored outside single one-off studies.

2.1.2 Trust and HRI Metrics

Trust is a key requirement for working with collaborative robots, as low levels of trust can lead to under-utilization in work and home environments [50]. In autonomous and robotic systems trust is generally agreed to be an positive attribute as well [51, 52]. A key component of the dynamic nature of trust is created in the first phase of a relationship [53, 54], while lack of early trust building can remove the opportunity for trust to develop later on [55]. Lack of trust in robotic systems can also lead to expert operators bypassing the robot to complete the task [56]. Trust is generally categorized into either cognitive trust and affective trust [57]. Perceiving emotion is a crucial for the development of affective trust in human-to-human interaction [58], as it increases the willingness to collaborate and expand resources bought to the interactions [59]. Importantly, relationships based on affective trust are more resilient to mistakes by either party [58], and perceiving an emotional identity has been shown to be an important contributor for creating believable and trustworthy interaction [4, 5]. In group interactions, emotional contagion has been shown to improve cooperation and trust in team exercises [60, 61, 62].

While trust is widely addressed in HRI literature, there are many other attributes with other metrics that have been develop to broaden understandings of HRI. One of the most widely used survey is the Godspeed Questionnaire Series, which measures anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots [63]. Each metric in the Godspeed survey is measured with 4-5 bipolar sub-questions. While the Godspeed metrics have recently shown to be problematic, we chose to use them for several reasons. One issue is the use of a bipolar-scale rating instead of a Likert scale [64]. We believe that having a high rating for Cronbach's Alpha somewhat alleviates the concern that each rating is not truly opposite, as that implies that at least each rating is internally reliable. In comparison to other alternate metrics that build on the Godspeed metrics, Robotic Social Attributes Scale (RoSAS) [65], the Godspeed was chosen as it allows us easy comparison between both our own existing studies, as well as many past studies that address the qualities from Godspeed with 2015 meta analysis indicating this survey had been used at least 160 times [66]. Other survey's are often created for a specific metric with more extended questions such as ratings for self-efficacy [67] or willingness to interact [68]. It is also very common for psychology and social studies metrics to be adopted within the field of robotics, such as the mind attribution scale [69].

2.1.3 Emotion

Emotion has received considerable attention in psychology with many different methods of classification emerging [70]. Most prominent amongst these is the discrete categorization by Ekman defining the six basic emotions, fear, anger, disgust, sadness, happiness and surprise [71]. Emotions are also often described through a continuous scale, with the most common being the circumplex model, mapping emotion on the two dimensions, valence and arousal [17]. A related term to emotion is mood, which is usually described as emotion over a longer time span, with more gradual shifts than emotion [72].

Emotion in robotics has seen dramatic increases in research over the last thirty years, spanning many applications and platforms [73]. This research can primarily be divided into two main categories, emotion for improved performance (called "survivability") and emotion for social interaction [74]. Survivability invokes the belief that emotion is key to animals' ability to survive and navigate the world and can likewise be used in robots. This includes situations such as an internal emotion based on external danger to a robot [75]. The second category - social interaction - addresses anyway emotion is used to improve interaction, such as analyzing a humans emotion, or portraying emotion to improve agent likeability and believability [76].

The ability to express emotion through nonverbal means can be an effective tool for computational and mechanical systems which interact with people. Coulson's component process view of emotions is defined as the "affective significance of a series of evaluations" [77]. These physical behaviors are usually an unconscious reaction and can be considered representations of one's internal states. There is a large body of evidence which supports facial expressions and prosodic cues as being indicative of a person's internal emotional state [7, 78, 79, 80]. It has also been shown that there is a relationship between external physical behavior and emotional states [81, 82], and evidence suggests that body language expresses emotional states better than facial expression [83]. Humans can associate postures and movements with particular emotions [84, 85, 86, 87, 88], while the human brain can process emotional body language without reliance on the primary visual cortex [89]. Gestural interaction is one of the important facets of emotional communication, as studies have shown that characteristics such as velocity, acceleration, and location can influence how people respond emotionally to robotic movement [90, 91]. In related work, it has been shown that robots which respond to humans using expressive behaviors enable more natural, engaging, and comfortable human-robotic interaction [92, 13, 93]. An emotionally responsive and expressive robot can, therefore, be effective in social situations such as in learning and teaching environments by communicating information such as compassion, awareness, accuracy, and competency through a non-conscious affective channel [94]. Lee [95] investigated how certain robotic gestures can lead to trust building with humans while other gestures might hamper trust in such social interactions. In group social settings, emotion conveying gestures can also be used to elicit responses from other contributors and improve group efficiency by minimizing social conflicts [96, 97, 98, 99].

One of the well known robots designed to express emotions is MIT's *Kismet*, which in addition to torso gestures can convey emotions using facial expression on a continuous valence, arousal, and stance scale [100]. Other efforts to design emotionally expressive robots include *Cathexis*, which is based on the six primary emotions of anger, fear, disgust, sadness, happiness and surprise [101], and *NAO* which uses automated scheduling of discrete motion primitives driven by beat and emotion in music [102]. A more recent system designed for human-machine companions integrates the use of both discrete and continuous emotions [103]. Here, external events are subjected to an appraisal process and interpreted as discrete emotional stimuli. A few related works have attempted to use robotic gestures to accompany speech to help convey affective information to the observer [104]. Gestures have also been found to serve as a component of the speech planning process, helping speakers to organize and conceptualize spatial information [105]. Informed by these systems, researchers such as Salem [106] have developed integrated models of speech-gesture production to address concurrence. However, to our knowledge, there is no prior work that attempts to integrate physical gestures and EMP to convey robotic emotional states, as proposed in this dissertation.

2.2 Computational Music

2.2.1 Emotion

While the underlying mechanisms of humans' emotional response to music are still being investigated and are under debate [107], it is agreed that music is a powerful medium for evoking emotions [14]. In the Music Information Retrieval (MIR) community, efforts have been made to understand the relation of emotion and music [108], focusing on Perceived Emotions (the emotions perceived by listeners to be expressed in the music), rather than Felt Emotions (the emotions felt by the listener while listening to music) [109]. In recent years, machine learning has become dominant for understanding emotions in music, and the MIREX audio mood Classification Task [110] has become a common base-line for categorical musical classification of emotions. Some efforts have been made to pre-train classifiers for alternative musical emotion categories such as passionate, cheerful, bittersweet, silly/quirky, and aggressive [111]. For modeling musical emotion based on the dimensional emotional model of valence and arousal, regression models such as Support Vector Regression (SVR) have been commonly used [112].

2.2.2 Human-Computer Interaction through Sound

Musical and audio-driven systems have been used to serve many areas across technological systems. Sonification, where audio is used to communicate data, has been widely integrated into computer systems. Examples include head gesture sonification to support social interaction for visually impaired persons [113]. Sonification has also been used to convey information for improving human-human interactions. Oh et al. sonified 2-d gestures in order to teach visually impaired persons to perform the gesture, finding that pitch best mapped to vertical movements while stereo panning best mapped to horizontal movements [114]. Many dance sonification projects have also been conducted, such as real-time audio based on dancers' movements [115]. Sonification of physical gestures has additionally been used widely in sports such as optimizing performance in rowing [116]. In robotics research, Zhang et al. studied robotic sonification in relation to emotions for children with Autism Spectrum Disorder (ASD) [48].

In other related work, sound has been used to increase trust in digital assistants through the addition of basic sound patterns [117]. In 1997, Alty et al. presented a position paper that described the potential of music to become a key communication medium for all technology [118]. They argue that audio is underused in many tasks such as code debugging. Other research using audio has focused on the potential for communicating graphical interfaces to the visually impaired [119]. Music has also been used for roles in technology such as a sports training guide to encourage casual runners [120] or to improve navigation while driving [121]. While these studies all suggest the vast potential for music in HCI and HRI, they are almost all single studies with systems that are used only for the experiment itself.

2.2.3 Generative Music

Computer generative music systems have been widely explored starting from systems in the 1950's [122]. Early systems in generative music can be primarily be split between rule-based systems or stochastic systems [123]. Rule-based systems focused on the creators musical choices, who often acted as programmer and composer, such as Padberg's Canon and Free Fugue which was inspired by 12-tone composition [124]. Stochastic systems often involve randomness and probability to generate music [125]. Musical generation has largely followed broader trends in computer science, with a higher emphasis on approaches based on methods from AI gradually emerging. Early uses of AI for music generation used processes such as Markov models, genetic algorithms and cellular automata, before gradually shifting to a high reliance on existing musical datasets [126].

Modern state of the art music generation systems primarily leverage deep learning [127]. These systems can focus on symbolic training, where the system learns from musical representations, such as MIDI. Music transformer, trained on classical piano music is able to generate music with long-term structure [128]. Many systems also focus on generating raw audio itself, Jukebox is able to generate full songs, sample-by-sample from audio [129]. Both the symbolic Music Transformer and the audio based Jukebox require extensive datasets, and extreme resources to train. They are also both only capable of generating offline, with no potential use in real-time systems. For real-time interactive systems, the primary focus is existing earlier approaches, such as Continuator which a collection of Markov models to generate phrases [130]. Overall, work for real-time computer musical phrase generation is significantly overlooked.

2.3 Prosody

Vocal prosody addresses the intonational and rhythmic aspects of spoken language that are not encoded by grammar or vocabulary, and bears strong resemblance to music in the manner it conveys emotions. Figure 2.1 shows the spectrogram for a single English phrase, spoken by the same actor in four emotions. This figure highlights the different pitch and timbre features between each emotion. In both music and prosody, emotions can be classified in a discrete categorical manner (happiness, sadness, fear, etc.) with more complex emotions considered as a combination of two or more of these fundamentals [131]. Prosody varies widely between languages and cultures, in tonal languages such as Mandarin, changing prosody features can convey different semantic meanings [132]. Other prosodic variations between languages include features such as the use of rhythm and timing between Spanish and English, while there exist wide variations in the use of pitch between English and Japanese [133].

Music shares many of the underlying building blocks of prosody such as pitch, timing, loudness, intonation, and timbre [10, 15, 16]. The relation between speech and music has been widely studied [134]. Research has categorized and compared the relationship between speech and music, with common acoustic features often employed and analyzed in the same manner [135, 136, 137].

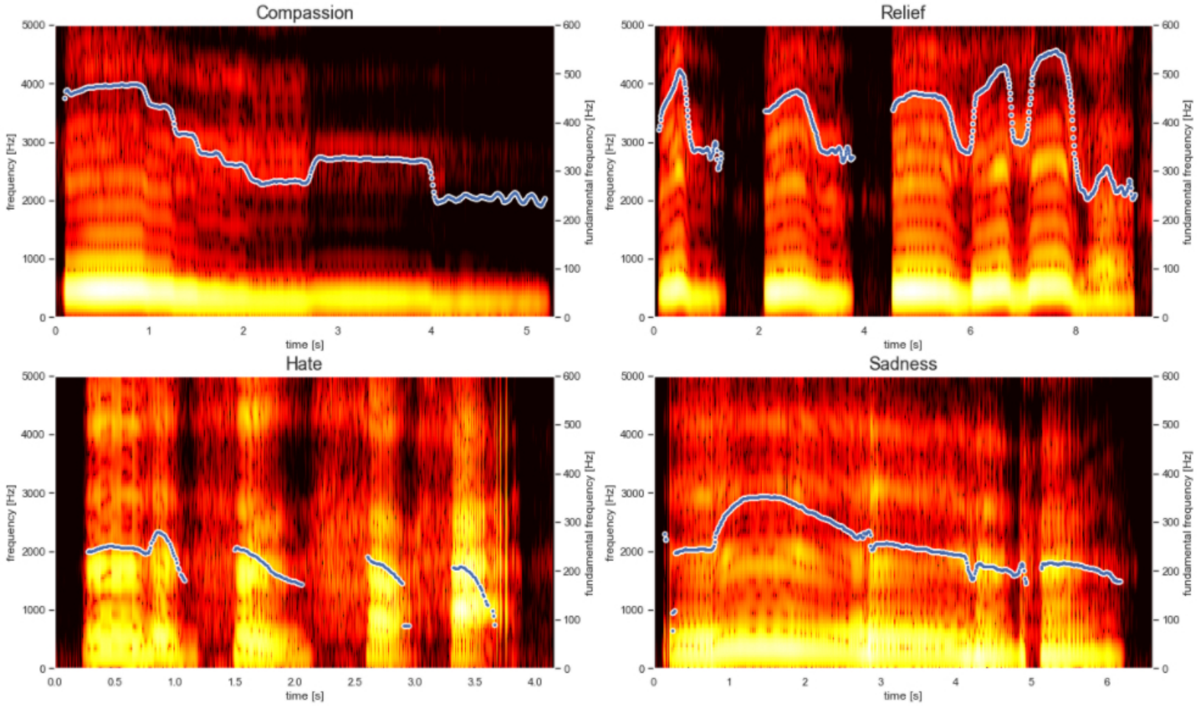


Figure 2.1: Spectrogram of prosodic phrases (blue line indicates pitch contour).

Prosody supporting language has been proven to be an effective communication channel to convey mood and emotions [138]. Robotic researchers have therefore attempted to model and manipulate emotion through prosody synthesis over the last 25 years [12].

Early rule-based prosodic systems such as in concatenative speech synthesizers [139] have been replaced by data-driven techniques such as Hidden Markov Models [140] and more recently Deep Learning [141], leading to significant advances in producing more natural sounding synthesized speech. While researchers in this field have explored both the categorical [142] and dimensional approaches [143] to modulate emotions little research has been done into how the emotional manipulation of robotic voice is perceived by human collaborators. While the basic elements of prosody such as pitch, timing, loudness, intonation, and timbre, are commonly used in music analysis and generation ([10], [15], [16]), no known efforts have been made to use models from music analysis to inform robotic prosody, as we propose here.

2.4 Summary

The research in this dissertation aims to develop new methods for robotic communication through the use of sound. To do this it combines literature and common research process from HRI, with generative software techniques and approaches from music technology. Broader approaches and understandings of prosody have been developed in both music technology and HRI, however prosody without language has been generally overlooked. We believe that by leveraging music technology processes to generate prosody we can develop new modes of interaction that can improve the field of HRI.

CHAPTER 3

GENERATING EMP FOR ROBOTICS

Text from this section has been published as:

Before, Between, and After: Enriching Robot Communication Surrounding Collaborative Creative Activities, Richard Savery, Lisa Zahray, Gil Weinberg, Frontiers in Robotics and AI: Creativity and Robotics, 2021 [144]

Musical Prosody-Driven Emotion Classification: Interpreting Vocalists Portrayal of Emotions Through Machine Learning, Nic Farris, Brian Model, Richard Savery, Gil Weinberg, 18th Sound and Music Computing Conference, 2021 [145]

Emotional Musical Prosody: Validated Vocal Dataset for Human Robot Interaction, Richard Savery, Lisa Zahray, Gil Weinberg The 2020 Joint Conference on AI Music Creativity (CSMC + MUME), 2020 [146]

To apply EMP in robotic systems, and develop new understandings of the impact on HRI, our first step was to develop a robust model for creating EMP. At the basis of our new model for EMP was the collection of a new Emotionally-Labeled-Musical-Prosody (ELMP) dataset. This dataset forms the foundation for generating EMP for robotics in real time. After collecting the dataset we conducted a range of classification tasks to validate the ability for machine learning features to separate the dataset by emotion. We then developed a new generative model trained on pitch sequences from the dataset combined with a scale-based audio sampler for real-time musical playback. Our first research question explores the use of the dataset for a deep learning generative system.

RQ 1: Can a data driven, EMP system generate musical phrases that can be labelled by listeners?

3.1 Dataset

Before collecting the data, we conducted exploratory sessions with seven different student musicians, comparing their ability to improvise different emotions using different classification systems. We additionally evaluated how well the musicians in this group could recognize the emotions played by other musicians. This process consisted of a 45 minute in-person session, with musicians first improvising, followed by an informal interview to discuss the difficulty and their preferences for emotional classifications for improvisation. After these sessions, we decided that the Geneva Emotion Wheel (GEW) [147] was best suited for our purposes. The GEW is a circular model, containing 20 emotions with emotions and position corresponding to the circumplex model.

Our decision to use the GEW was based on multiple factors, firstly we aimed to capture as large a range of emotions as possible, that could be accurately improvised by musicians in the sessions. In our exploratory study, the GEW balanced between having many recognizable classes, while also avoiding the potential confusion from too many overlapping classes, or the challenge of continuous classes such as the circumplex model. The GEW also has advantages for implementation, with 20 different discrete emotions which can be reduced to four separate classes, aligned with a quadrant from the circumplex model. GEW also includes most of Ekman's basic emotions - fear, anger, disgust, sadness, happiness - only leaving out surprise. The ability to potentially reduce our collected dataset between these different models of emotion allows for significant future use cases.

We first created a short list of vocalists who we have worked with in the past. We then conducted Skype calls with multiple professional vocalists refining the overall plan and describing the process. The final three vocalists that recorded phrases for the dataset were Mary Esther Carter [1], Ella Meir [2] and Aya Inohue. All three are professional vocalists and improvisers who the authors have worked with before, and were confident would be able to

[1] https://maryesthercarter.com/

[2] https://www.ellajoymeir.com/

create a dataset matching the projects goals. Additionally, each vocalist has a performance and acting background, which led us to believe that they would be able to not just sing emotionally, but have a higher perception of how emotion would appear to a broad external audience.

Each vocalist had at home access to high quality recording equipment. The vocalists were paid $500 to record the samples over a week long period at her home studio, using a template we created in Apple digital audio workstation - Logic Pro, while maintaining the same microphone positioning. For the samples we requested phrases to be between 1 and 20 seconds, and to spend about 15 minutes on each emotion, allowing unscripted jumping between any order of the emotions. We allowed deletion of a phrase if the singer felt retroactively that the phrase did not capture the correct emotion. The final recorded dataset includes 8863 phrases equalling 14 hours of data with an average of 443 phrases for each emotion. Samples from the dataset can be heard online.[3]

It should be noted that this dataset comes from three musicians, and therefore only captures three individuals perspectives on musical emotion. While the dataset can make no claim to represent cross-cultural emotion conveyance and does not create a generalized emotion model, we believe that only collecting data from three musicians has advantages. By having vocalists of similar style our system can recreate a general emotional style, avoiding incorrectly aggregating multiple styles to remove distinctive individual and stylistic features.

Dataset Validation

To validate the dataset, we conducted a study with 45 participants for each vocalist from Prolific and Mechanical Turk, paying each $3 for 10 minutes. Each question in the survey asked the participant to listen to a phrase and select a location on the wheel corresponding to the emotion and intensity they believed the phrase was trying to convey. Phrases fell

[3] www.richardsavery.com/prosodycvae

under two categories of “best” and “all”, with each participant listening to 60 total phrases selected at random. Between the 45 participants listening to 60 phrases, 2700 ratings were given per vocalist, which we believe gave a strong overall rating of the dataset.

The “all” category consisted of all phrases in the dataset for that emotion, with a new phrase randomly selected for each participant. The best emotions were chosen to ensure an even distribution of phrase lengths in each emotion set, with each emotion having a chosen phrase for the lengths, 3, 5, 7, 9, and 11 seconds. When multiple phrases existed for each length the authors chose phrases that were most distinctive in style from the other emotions, aiming to create a stylistic separation between each emotion class.

To validate the data we primarily compared participants responses to their ability to recognize the emotion by quadrant. We focused on comparing by quadrant as we believed for the purpose of future studies participants being able to in-distinguish between categories such as “love” and “admiration” was not as critical. We additionally marked phrases that had an accuracy of over 80% for future use, which were most common for disgust and sadness.

We computed the mean and variance for each emotion, weighted by intensity, using the methods described in [148], which rely on circular statistics. The results are shown in Table 3.1, with a comparison to our generated phrases which are discussed in Section 3.2. The first columns show the percentage of all data points that were classified as an emotion in the correct quadrant. The next columns, showing average difference, were calculated by first finding the difference between each ground truth emotion’s angle and its weighted average reported angle, and then averaging that value over the emotions within each quadrant. It is worth noting that only three emotions in the dataset and two emotions in the generated data had weighted average angles outside the correct quadrant. The final units were converted from degrees to units of emotion (20 emotions in 360 degrees). The last columns, showing variance, were calculated by finding the weighted variance for each emotion (converted to units of emotion), and then averaging for each quadrant.

Table 3.1: Results of emotion survey for dataset phrases compared with generated phrases. See Section 3.2.3 for an explanation of the metrics.

Quadrant	**% Correct Quadrant**		**Average Difference**		**Average Variance**	
	Dataset	**Generated**	**Dataset**	**Generated**	**Dataset**	**Generated**
1	57.2	56.3	1.32	1.98	1.76	1.83
2	54.5	52.5	1.45	0.96	1.79	1.88
3	57.4	51.5	2.16	1.93	1.92	1.89
4	43.7	31.9	1.61	1.24	1.86	2.03

Table 3.2: Features Extracted

ID	Feature
1	Zero Crossing Rate
2	Energy
3	Entropy of Energy
4	Spectral Centriod
5	Spectral Spread
6	Spectral Entropy
7	Spectral Flux
8	Spectral Rolloff
9-21	MFCCs
22-33	Chroma Vector
34	Chroma Deviation

3.1.1 Dataset Analysis

To demonstrate the potential of our dataset we conducted multiple machine learning experiments. Prior work focusing on emotion music classification has found success in the implementation of k-nearest neighbor (K-NN) and support vector machines (SVM), finding the highest accuracies using SVMs [149]. In exploration of the relationship of feature extraction techniques and their contribution toward emotional classification, we implemented a variety of machine learning models, and trained and evaluated KNNs, linear SVMs, Random Forests, Extra Trees, Gradient Boosting, and Feed Forward Neural Networks (FFNN).

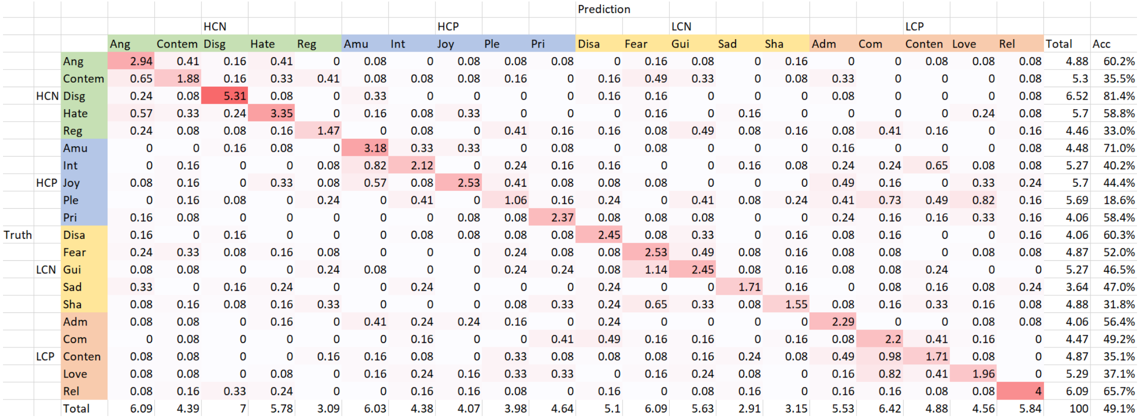
We first analyzed Mary Carter's dataset alone, to attempt to categorize emotion for one singer before generalizing across all three. We chose to start with only one vocalist to be sure we could classify the vocalists individually, before generalizing acroos all three. We analyzed the baseline accuracies, F-scores, and confusion matrices achieve by training KNNs, linear SVMs, Random Forests, Extra Trees, Gradient Boosting models utilizing

Table 3.3: Single Taxonomy, 1 Singer Classification Results

Model	Accuracy	F1	Hyperparam
KNN	33.8	32.1	C=15
SVM	49.1	48.1	C=5.0
Extra Trees	44.3	42.8	C=500
Gradient Boosting	47.2	46.6	C=200
Random Forest	43.8	42.3	C=200

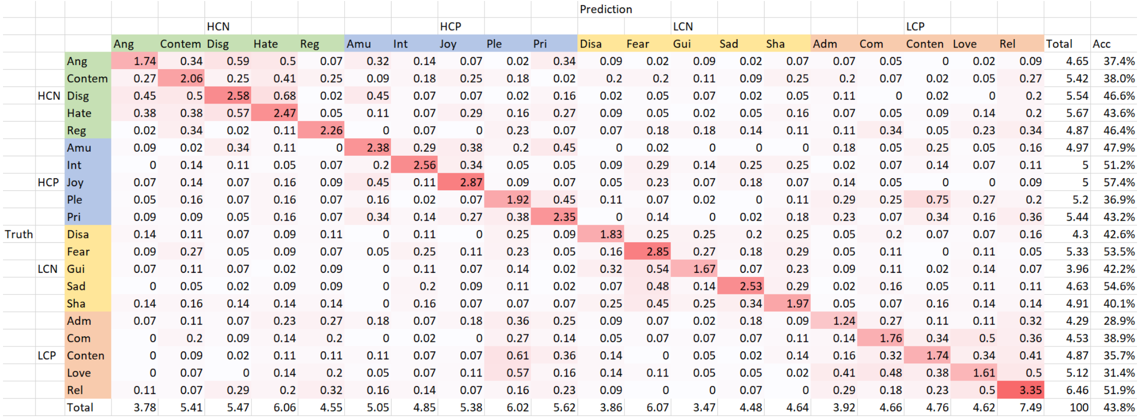
the audio features outlined in Table 3.2. Table 3.3 shows the best accuracy, F1-score, and selected hyper-parameter for each of our models. Each model significantly outperforms random guessing. Even the worst model, the KNN, performs 6.5 times better than random chance (20 possible categories = 5% chance random guessing). Our best model, the linear SVM, performs approximately 10 times better than random guessing with an accuracy of 49.1%. The confusion matrix for the single emotion taxonomy has been included in Figure 3.1. Analysis of this confusion matrix yields a few observations: Disgust is rarely confused with other emotions, having the highest individual accuracy of 81.4%. We expect that disgust is easily categorized as in our review of the dataset the vocalists often used specific timbres to emphasis the emotion. Disgust has also shown significant similarities in human's vocalization across English speaking groups [150]. Fear and Guilt are the two most common pair of emotions to be confused for one another. Pleasure is the most difficult emotion for the model to classify correctly, having the lowest individual accuracy of 18.6%. Our models also perform extremely well when tasked with categorizing between two emotions, achieving accuracies as high as 98.9% with a f1 of 98.9 in the distinction between Love and Disgust using a SVM. This reinforces the intuition that by reducing the number of emotional categories we can achieve higher accuracies for identification.

We next evaluated the ability for these models to generalize across all three singers. With the exception of linear SVM, all model architectures maintain similar accuracies when trained on the 3 singer datasets, showing the ability of emotion to generalize across this dataset. Figure 3.2 shows the confusion matrix for all three singers.



												Prediction											
					HCN					HCP					LCN					LCP				
			Ang	Contem	Disg	Hate	Reg	Amu	Int	Joy	Ple	Pri	Disa	Fear	Gui	Sad	Sha	Adm	Com	Conten	Love	Rel	Total	Acc
		Ang	2.94	0.41	0.16	0.41	0	0.08	0	0.08	0.08	0.08	0	0.16	0.08	0	0.16	0	0	0.08	0.08	0.08	4.88	60.2%
		Contem	0.65	1.88	0.16	0.33	0.41	0.08	0.08	0.08	0.16	0	0.16	0.49	0.33	0	0.08	0.33	0	0	0	0.08	5.3	35.5%
	HCN	Disg	0.24	0.08	5.31	0.08	0	0.33	0	0	0	0	0.16	0.16	0	0	0	0.08	0	0	0	0.08	6.52	81.4%
		Hate	0.57	0.33	0.24	3.35	0	0.16	0.08	0.33	0	0	0	0.16	0	0.16	0	0	0	0	0.24	0.08	5.7	58.8%
		Reg	0.24	0.08	0.08	0.16	1.47	0	0.08	0	0.41	0.16	0.16	0.08	0.49	0.08	0.16	0.08	0.41	0.16	0	0.16	4.46	33.0%
		Amu	0	0	0.16	0.08	0	3.18	0.33	0.33	0	0.08	0	0.08	0	0	0	0.16	0	0	0	0.08	4.48	71.0%
		Int	0	0.16	0	0	0.08	0.82	2.12	0	0.24	0.16	0.16	0	0	0.16	0.08	0.24	0.24	0.65	0.08	0.08	5.27	40.2%
	HCP	Joy	0.08	0.16	0	0.33	0.08	0.57	0.08	2.53	0.41	0.08	0.08	0.08	0	0	0	0.49	0.16	0	0.33	0.24	5.7	44.4%
		Ple	0	0.16	0.08	0	0.24	0	0.41	0	1.06	0.16	0.24	0	0.41	0.08	0.24	0.41	0.73	0.49	0.82	0.16	5.69	18.6%
		Pri	0.16	0.08	0	0	0	0	0	0.08	0.08	2.37	0.08	0.08	0.08	0	0	0.24	0.16	0.16	0.33	0.16	4.06	58.4%
Truth		Disa	0.16	0	0.16	0	0	0	0.08	0.08	0.08	0.08	2.45	0.08	0.33	0	0.16	0.08	0.16	0	0	0.16	4.06	60.3%
		Fear	0.24	0.33	0.08	0.16	0.08	0	0	0	0.24	0.08	0.08	2.53	0.49	0.08	0.16	0.08	0.08	0	0.08	0.08	4.87	52.0%
	LCN	Gui	0.08	0.08	0	0	0.24	0.08	0	0	0.24	0.24	0.08	1.14	2.45	0.08	0.16	0.08	0.08	0.24	0	0	5.27	46.5%
		Sad	0.33	0	0.16	0.24	0	0	0.24	0	0	0	0.24	0	0	1.71	0.16	0	0.08	0.16	0.08	0.24	3.64	47.0%
		Sha	0.08	0.16	0.08	0.16	0.33	0	0	0	0.08	0.33	0.24	0.65	0.33	0.08	1.55	0.08	0.16	0.33	0.16	0.08	4.88	31.8%
		Adm	0.08	0.08	0	0.16	0	0.41	0.24	0.24	0.16	0	0.24	0	0	0	0	2.29	0	0	0.08	0.08	4.06	56.4%
		Com	0	0.08	0	0	0	0	0.16	0	0	0.41	0.49	0.16	0.16	0	0.16	0.08	2.2	0.41	0.16	0	4.47	49.2%
	LCP	Conten	0.08	0.08	0	0	0.16	0.16	0.08	0	0.33	0.08	0.08	0.08	0.16	0.24	0.08	0.49	0.98	1.71	0.08	0	4.87	35.1%
		Love	0.08	0.08	0	0.08	0	0.16	0.24	0.16	0.33	0.33	0	0.16	0.24	0.08	0	0.16	0.82	0.41	1.96	0	5.29	37.1%
		Rel	0.08	0.16	0.33	0.24	0	0	0.16	0.16	0.08	0	0.16	0	0.08	0.16	0	0.16	0.16	0.08	0.08	4	6.09	65.7%
		Total	6.09	4.39	7	5.78	3.09	6.03	4.38	4.07	3.98	4.64	5.1	6.09	5.63	2.91	3.15	5.53	6.42	4.88	4.56	5.84	100	49.1%

Figure 3.1: SVM, Individual Taxonomy, 1 Singers Confusion Matrix



												Prediction											
					HCN					HCP					LCN					LCP				
			Ang	Contem	Disg	Hate	Reg	Amu	Int	Joy	Ple	Pri	Disa	Fear	Gui	Sad	Sha	Adm	Com	Conten	Love	Rel	Total	Acc
		Ang	1.74	0.34	0.59	0.5	0.07	0.32	0.14	0.07	0.02	0.34	0.09	0.02	0.09	0.02	0.07	0.07	0.05	0	0.02	0.09	4.65	37.4%
		Contem	0.27	2.06	0.25	0.41	0.25	0.09	0.18	0.25	0.18	0.02	0.2	0.2	0.11	0.09	0.25	0.2	0.07	0.02	0.05	0.27	5.42	38.0%
	HCN	Disg	0.45	0.5	2.58	0.68	0.02	0.45	0.07	0.07	0.02	0.16	0.02	0.05	0.07	0.02	0.05	0.11	0	0.02	0	0.2	5.54	46.6%
		Hate	0.38	0.38	0.57	2.47	0.05	0.11	0.07	0.29	0.16	0.27	0.09	0.05	0.02	0.05	0.16	0.07	0.05	0.09	0.14	0.2	5.67	43.6%
		Reg	0.02	0.34	0.02	0.11	2.26	0	0.07	0	0.23	0.07	0.07	0.18	0.18	0.14	0.11	0.11	0.34	0.05	0.23	0.34	4.87	46.4%
		Amu	0.09	0.02	0.34	0.11	0	2.38	0.29	0.38	0.2	0.45	0	0.02	0	0	0	0.18	0.05	0.25	0.05	0.16	4.97	47.9%
		Int	0	0.14	0.11	0.05	0.07	0.2	2.56	0.34	0.05	0.05	0.09	0.29	0.14	0.25	0.25	0.02	0.07	0.14	0.07	0.11	5	51.2%
	HCP	Joy	0.07	0.14	0.07	0.16	0.09	0.45	0.11	2.87	0.09	0.07	0.05	0.23	0.07	0.18	0.07	0.14	0.05	0	0	0.09	5	57.4%
		Ple	0.05	0.16	0.07	0.16	0.07	0.16	0.02	0.07	1.92	0.45	0.11	0.07	0.02	0	0.11	0.29	0.25	0.75	0.27	0.2	5.2	36.9%
		Pri	0.09	0.09	0.05	0.16	0.07	0.34	0.14	0.27	0.38	2.35	0	0.14	0	0.02	0.18	0.23	0.07	0.34	0.16	0.36	5.44	43.2%
Truth		Disa	0.14	0.11	0.07	0.09	0.11	0	0.11	0	0.25	0.09	1.83	0.25	0.25	0.2	0.25	0.05	0.2	0.07	0.07	0.16	4.3	42.6%
		Fear	0.09	0.27	0.05	0.09	0.07	0.05	0.25	0.11	0.23	0.05	0.16	2.85	0.27	0.18	0.29	0.05	0.11	0	0.11	0.05	5.33	53.5%
	LCN	Gui	0.07	0.11	0.07	0.02	0.09	0	0.11	0.07	0.14	0.02	0.32	0.54	1.67	0.07	0.23	0.09	0.11	0.02	0.14	0.07	3.96	42.2%
		Sad	0	0.05	0.02	0.09	0.09	0	0.2	0.09	0.11	0.02	0.07	0.48	0.14	2.53	0.29	0.02	0.16	0.05	0.11	0.11	4.63	54.6%
		Sha	0.14	0.16	0.14	0.14	0.14	0	0.16	0.07	0.07	0.07	0.25	0.45	0.25	0.34	1.97	0.05	0.07	0.16	0.14	0.14	4.91	40.1%
		Adm	0.07	0.11	0.07	0.23	0.27	0.18	0.07	0.18	0.36	0.25	0.09	0.07	0.02	0.18	0.09	1.24	0.27	0.11	0.11	0.32	4.29	28.9%
		Com	0	0.2	0.09	0.14	0.2	0	0.02	0	0.27	0.14	0.05	0.07	0.07	0.07	0.11	0.14	1.76	0.34	0.5	0.36	4.53	38.9%
	LCP	Conten	0	0.09	0.02	0.11	0.11	0.11	0.07	0.07	0.61	0.36	0.14	0	0.05	0.02	0.14	0.16	0.32	1.74	0.34	0.41	4.87	35.7%
		Love	0	0.07	0	0.14	0.2	0.05	0.07	0.11	0.57	0.16	0.14	0.11	0.05	0.05	0.02	0.41	0.48	0.38	1.61	0.5	5.12	31.4%
		Rel	0.11	0.07	0.29	0.2	0.32	0.16	0.14	0.07	0.16	0.23	0.09	0	0	0.07	0	0.29	0.18	0.23	0.5	3.35	6.46	51.9%
		Total	3.78	5.41	5.47	6.06	4.55	5.05	4.85	5.38	6.02	5.62	3.86	6.07	3.47	4.48	4.64	3.92	4.66	4.76	4.62	7.49	100	43.8%

Figure 3.2: Gradient Boosting, Individual Taxonomy, 3 Singers Confusion Matrix

3.2 Symbolic Generation

The generative system was designed with the primary goal of operating and responding to audio in real time on multiple robotic platforms. In past work we have generated raw-audio for EMP [151], however even after considerable refinement, and the use of multi-GPU systems, generation required three seconds of processing per one second of audio. With this in mind the initial design choice was to generate symbolic data using a version of the dataset converted to MIDI values, and not attempt to generate raw audio.

The symbolic generation of the system contains the pitch and rhythm of emotionally labelled melodies. Due to the process described in Section 3.2 the data also includes micro-timings. Symbolic data alone does not capture the range of emotion present in the dataset through timbre variations. By using the scale dataset described in Section 3.2 the generation process encapsulates symbolic information with tagged emotion, capturing timbre and phoneme information. Figure 3.4 shows an overview of the system. The system's interface is written MaxMSP, allowing users to chose an emotion. This activates a python script which generates a midi file and returns it to MaxMSP. Generated samples can be heard online.[4]

Dataset to MIDI

We converted each phrase's audio into a midi representation to use as training data. This was done to allow symbolic training on the dataset, instead of using only the raw audio. This process required significant iteration, as we developed a custom pipeline for processing our dataset. This was necessary due to the range of vocal timbre and effect, ranging from clear melodies, to non-pitched effects. We first ran the monophonic pitch detection algorithm CREPE [152] on each phrase, which output a frequency and a confidence value for a pitch being present every 0.01 seconds. As the phrases included breaths and silence, it was necessary to filter out pitches detected with low confidence. We applied a threshold followed by a median filter to the confidence values. We next converted the frequencies to midi pitches. We found the most common pitch deviation for each phrase using a histogram of deviations, shifting the midi pitches by this deviation to tune each phrase. This process allowed us to have a primary note at all times, while maintaining the sampled vibrato at 0.01 seconds. We rated onsets timing confidence between 0 and 1. To address glissando, vibrato and other continuous pitch changes, we identified peaks in the absolute value of the pitch derivative, counting an onset only when detecting a pitch for at least 0.04 seconds.

In converting the dataset to MIDI a primary concern was to maintain as many features from the dataset as possible. By sampling at 0.01s we were able to capture many pitch features in addition to the main melody, including techniques such as vibrato and glissando's.

[4] www.richardsavery.com/prosodycvae

This resolution of sampling also allowed us to capture micro-timings, and subtle rhythmic variation that would not be possible at a higher rate. We also sampled volume levels at 0.01s intervals allowing us to maintain variation in dynamics for each phrase.

Audio Sampler

In addition to the primary data collection of EMP we asked the vocalists to record chromatic scales at a range of tempos. This was done to allow us a way to playback the newly generated phrases while capturing as much vocal prosody from the sampled notes as possible. Our goal was that by combining the pitch and rhythm features captured from our dataset to MIDI process, with the timbre features we would capture as many prosodic features as possible, while still being capable of real-time generation. The data collection plan was based around common practice described by virtual instrument libraries [5]. For each emotion, 11 versions of a chromatic scale across an octave and a half were sung, 3 with very short notes, 3 with 500ms, 3 with 1000ms and 2 with 2000ms duration. To allow the scales to contain all timbrel features for each emotion, we allowed for any dynamic variations and accents. The syllables themselves were also chosen for each scale by the vocalist.

Scherer has shown that musical scales - without a melody or rhythm - are able to display emotion [153]. We therefore asked the singer to also record scales tagged with emotion to be used in an audio sampler. The audio sampler was designed to play back each note from the recorded scales, in such a way that new symbolic phrases consist of combinations of each note from the scale. In contrast to the main dataset we only recorded scales for four emotion classes, corresponding with each quadrant of the circumplex model. In addition to explaining the model to the vocalist, each quadrant had two key words which were angry/anxious, happy/exciting, relaxing/serene, sad/bored.

[5] https://www.spitfireaudio.com/editorial/in-depth/grow-your-own-samples/

3.2.1 CC-VAE

Data Representation

We maintain the same data structure as developed in our audio to midi process, using midi pitch values that are sampled every 10 milliseconds. We then convert each melody to a length of 1536 samples, and zero pad shorter melodies. Versions of each phrase are then transposed up and down six semitones, to give 12 versions of each phrase, one in each key. The melody is then reshaped to be 32 by 48 samples. The emotion label for each melody is converted to a one-hot representation.

Network Design

We chose to use VAEs due to their recent success in sequence and music generation tasks, and because they allow for analysis of the latent space which can provide insight into how well the network has learned to represent the different emotions. VAEs can be used to generate new data by sampling and decoding from the latent space, allowing the system to learn features of the data in an unsupervised manner. Figure 3.3 shows the latent space after training a Vanilla VAE on our custom dataset, without emotion labels. This demonstrates the latent space is able to separate by emotion without conditioning.

Our Conditional VAE is based on the standard architecture proposed by Sohn et al. [154]. A Conditional Variational Encoder (CVAE) varies from a VAE by allowing an extra input to the encoder and decoder. We input a one-hot emotion label, allowing for sampling a specific emotion from the latent space. As is common practice for a VAE, we use Kullback-Leibler divergence in the loss function. Our latent space dimension is 512, which we arrived at after testing multiple variations.

We chose to use a Convolutional Network (ConvNet) within our CVAE for multiple reasons. Although ConvNets are much less common in symbolic music generation [127], they have been used for audio generation such as WaveNet [155] as well as some sym-

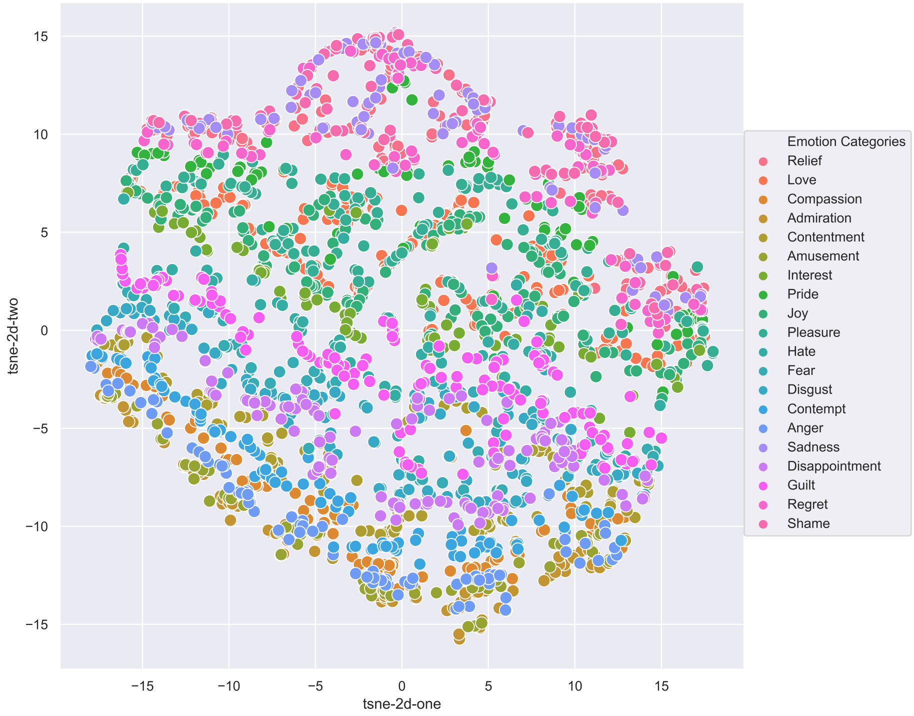


Figure 3.3: Vanilla VAE Latent Space, classifying Carter's audio dataset.

bolic generations [156]. While we experimented with Vanilla RNNs, LSTMs and GRUs as encoders and decoders we found they were very prone to overfitting when trained conditionally, likely due to our dataset size. Our architecture is presented in Figure 3.4.

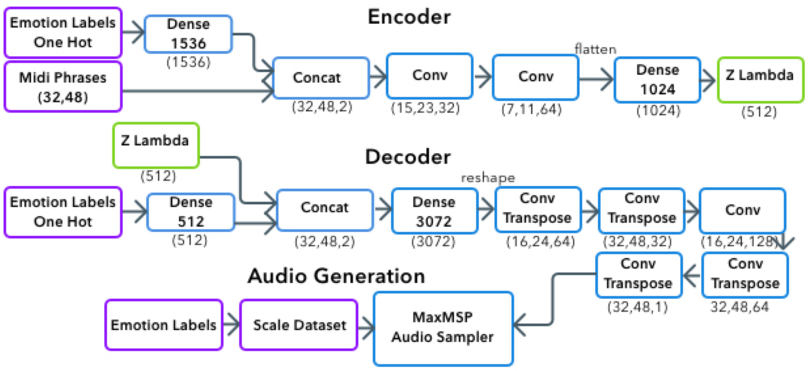


Figure 3.4: Generative System Overview.

3.2.2 Sample Player

The generated midi file is loaded into MaxMSP to be played by the sampler. The audio sampler plays back individual notes created during the recording of the scales. MaxMSP parses the midi file, assigning each note a midi channel. Channels are divided by emotion and note length. For example, happy is assigned to channels 1-4, with channel 1 containing the shortest note and channel 4 the longest note; sad is assigned to channels 5-8 with the shortest note assigned to channel 5 and the longest note assigned to channel 8. The audio sampler plays as a midi device, and can be played directly like any midi instrument.

3.2.3 Generation Evaluation

To evaluate the results, we first generated three phrases for each emotion. We then ran a survey using the same questions as the dataset validation described in Section 3.1, asking 39 new participants to select an emotion and intensity for each of the 60 total generated phrases. Participants encountered five attention checks during the survey, asking them in spoken word to chose a specific emotion, and we only used data from participants who answered all listening tests correctly. Figure 3.5 shows a comparison between the rose plots for each quadrant of the original dataset versus the generated phrases. Table 3.1 shows a direct comparison between the results for generated audio and the original dataset.

Rose plots of the validation results that combine the “best” and “all” categories can be seen in Figure 3.5, separated into each Geneva Wheel quadrant. The rose plot compares the collected dataset with the validation of our generated phrases. The plots show strong validation correlation in Quadrants 1, 2 and 3, while Quadrant 4 showed a higher level of confusion.

Our results show that the generated phrases performed similarly to the dataset in terms of emotion classification. While the percentage of phrases identified in the correct quadrant is slightly lower for the generated phrases, the average difference and variance have similar values. Visually, the rose plots show that participants were able to largely identify the

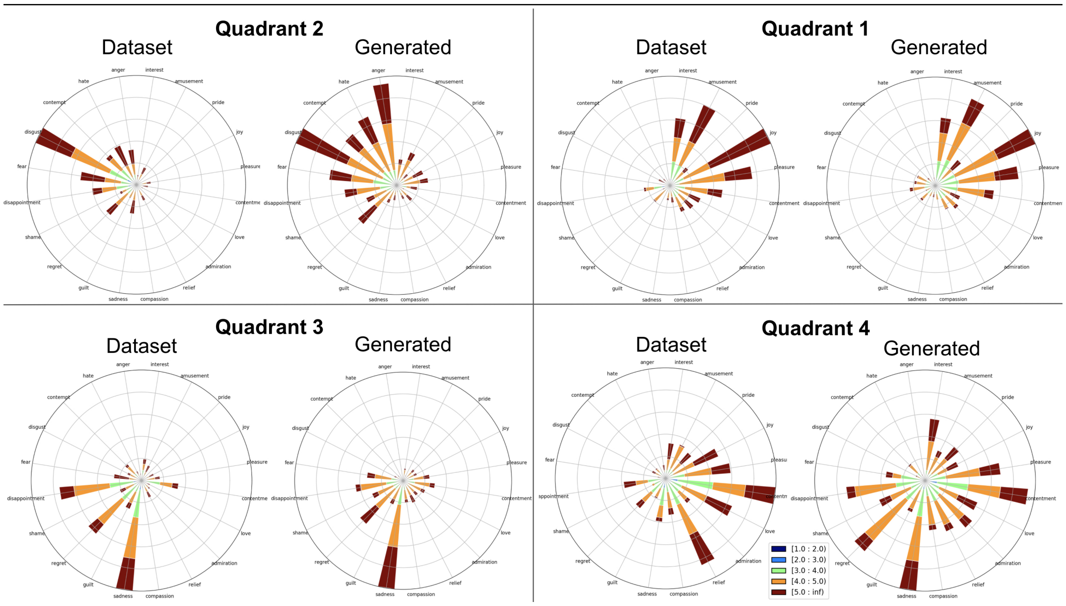


Figure 3.5: Rose plots of dataset validation and generation evaluation.

correct quadrant, having the most difficulty with Quadrant 4 (relaxing/serene) for both our collected dataset and generations.

Discussion

Our overall accuracy presented in Table 3.1 shows consistent results in the mid 50%. We believe this accuracy is acceptable for our current system, as the average variance and average difference are both close to two across all categories, implying that the primary errors in identification where small, such as mistaking love for admiration. For our the future experiments described in this dissertation we will use only specific generated EMP that score over 80% accuracy.

In both the original dataset and generated material participants had the lowest accuracy identifying the fourth quadrant emotions. Our results are not easily compared to other generative systems as the fourth quadrant emotions are rarely used in robotic studies [157]. This is partly because common classification systems such as Ekman's discrete classes do not include anything in the fourth quadrant. We also believe these emotions tend to be less easily displayed externally as they are low arousal and closer to neutral emotions. In future work we aim to consider methods to better develop the fourth quadrant emotions.

Our dataset used interpretation of emotions from one vocalist. While this had the benefit of consistency throughout phrases, in future work we intend to gather data from a larger number of musicians and to evaluate how well the model can generalize. We also plan to have other robots communicating through EMP using data from different vocalists.

3.3 Conclusion

In this chapter we presented a newly created dataset for EMP. We were able to accurately classify the dataset with machine learning. Our studies also showed that human listeners were able to label the emotions with high accuracy when considered across the four quadrants. We were then able to generate new phrases using a combination of a symbolic generations system and an audio sampler. We found the new generated phrases performed similarly in listening tests to the collected dataset.

CHAPTER 4
THE EFFECT OF EMBEDDING EMP IN DIFFERENT ROBOTIC PLATFORMS

Continuing from the development of an EMP generator, the next research question aims to analyze multiple potential use cases of EMP in individual robots. To this end we developed four separate studies to analyze EMP across four robotic platforms. For each platform we used the generated phrases described in the previous chapter, always choosing phrases that had been validated with an accuracy of over %80. Comparing common metrics across each platform led to the second research question:

RQ 2: How does EMP alter the level of likeability, perceived intelligence and trust in social, industrial, humanoid and robotic musicians?

We compared multiple robotic platforms as we believed the impact of EMP will vary by robot type and the type of interaction. Conducting multiple studies also allowed us to compare EMP to different audio types as appropriate for the robot, such as speech or simple non-emotion driven audio variations.For the co-bot industrial arm we conducted a study with the robot function as a collaborator for a pattern recognition task. For this task we recorded participant ratings for anthropomorphism, animacy, likeability, perceived intelligence and trust, each of which has been shown as a significant contributor for collaborative interaction and analyzed extensively [66, 158]. We replicated this pattern recognition task with a humanoid robot. For the social robot we programmed Shimi to engage in an emotion based exchange. We only measured trust with the social robot, allowing us to use a longer survey method. For studying a musical robot, we used Shimon the marimba playing robot with the metrics analyzed anthropomorphism, animacy, likeability and perceived intelligence. For the experiments with Shimon, we also analyzed whether using EMP alters the perception of Shimon's musical ability itself.

The robot gestures in each experiment were hand-designed following a set of exisiting

Table 4.1: Summary of Chapter 4 Experiments

Ch.	Robot	DOF	Metrics	Interaction Type	Comparing
4	Arm (Simulation)	4	Anthropomorphism, Intelligence, Likeability, Trust	Pattern recognition	EMP, gestures, non-prosody audio
4	Humanoid (Simulation)	18	Anthropomorphism, Intelligence, Likeability, Trust	Pattern recognition	EMP, gestures, non-prosody audio
4	Social Robot (Shimi)	5	Trust	Social Interaction	EMP, Speech
4	Robotic Musician (Shimon)	5	Creativity (Coherence, Novelty, Expressivity), Animacy, Anthropomorphism, Likeability, Intelligence	Musical improvisation	EMP, Speech

guidelines for emotional gestures [159]. This allowed the gestures to be mapped directly to and mapped around the generated EMP, and cater for the different degrees of freedom and types of movement available to each robot platform.

A summary of the robots and metrics analyzed is shown in Table 4.

4.1 EMP for Industrial and Humanoid Robotics

Text from this section has been published as:

EMP for the Enhancement of Trust in Robotic Arm Communication, Richard Savery, Lisa Zahray, Gil Weinberg, Trust, Acceptance and Social Cues in Human-Robot Interaction, Ro-MAN 2020 [160]

Industrial co-robotic arms are showing a significant expansion in use, which is expected to continue and grow into the foreseeable future [161]. While the use of such robotic arms expands, they still lack a standard form of non verbal communication [31]. Many non verbal methods to establish communication between robot arms and humans, such as haptics [162] or mixed reality [163], are costly to implement from a technical and financial perspective requiring custom equipment and training. More recent research has shown the importance of social and emotion communication for robots [164]. For industrial collaborative robots, displaying emotion has been shown to increase key metrics, such as the likelihood of humans to follow social norms[165], supporting better engagement with disability [166] and improving the perception of the robot as an equal human collaborator [167].

We believe EMP is uniquely positioned for industrial arms as it can enhance social interaction and engagement with human collaborators without requiring a change in core functionality. While robotic arms themselves do not generally approach the uncanny valley, we believe that independent modalities (such as a human voice) can cause the same impact on an interaction. This extends the notion of the habitability gap [1], where issues with interaction occur when a robot's functionality does not match its capability. In a robotic arm this could occur when a simple task, such as repeatedly moving an object, is accompanied by rich language based communication method. We firstly evaluate these interactions to confirm that there is no impact through potential distraction in collaboration with a robotic arm. We then measure how EMP compares to single-pitch audio and no audio systems for establishing trust, trust recovery, likeability and the perception of intelligence and safety. Finally, we analyze the same EMP on a humanoid robot to understand if and how our generative system can be transferred across systems and generalized.

4.1.1 Experiment

We conducted two different studies, one using a robotic arm and the other using a humanoid robot in an effort to address the following research questions.

Research Questions and Hypotheses

Our first research question focuses on understanding the role of EMP and trust in robotic interaction.

RQ1 *How does EMP alter trust and trust recovery from mistakes, compared to control conditions of no audio and single-pitch audio?*

For this question our hypothesis is that the overall trust at the end of the interaction will be significantly higher for EMP over single-pitch and higher for single-pitch audio over no audio. Our second research question compares common HRI metrics such as the perceived intelligence, perceived safety and likeability, for each robotic system.

RQ2 *How does EMP alter perceived safety, perceived intelligence and likeability?*

For the first two research questions, we believe that participants will develop an internal model of the robot as an interactive emotional collaborator for the EMP model. This will lead to higher levels of trust and improved perception of safety and intelligence. The third question explores the relation between users' self-reported metrics, gathered through surveys and their actual responses collected through a performance based task. We are interested in comparing whether the system that is self-reported to be trusted is actually utilized more in performance based tasks.

RQ3 *When a user self-reports higher levels of trust in a robot, does this in turn lead to higher utilization and trust in a robotic arm's suggestions?*

For this questions we hypothesize that users' self-reported trust ratings will correspond to their actual use and trust levels, as implied by choice to follow the decisions of the robotic system. We also hypothesize that by utilizing EMP, human collaborators will be more likely to trust the robotic arm's suggestions directly after a mistake.

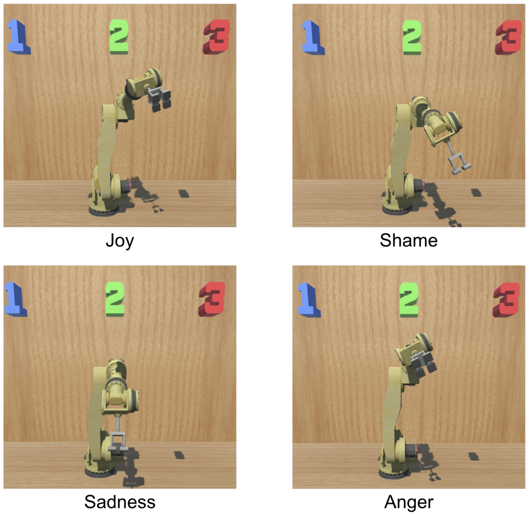
Experimental Groups and Robot Reactions

Our study was designed as a between-group experiment, where participants were randomly allocated to one of three conditions. These conditions were a EMP group, a single-pitch audio group (notes), and a control with no audio (gesture). In all three versions of the experiment, the robot responded with the emotional gestures.

In the EMP group, the gestural response was accompanied by playing an EMP phrase, randomly selected each time from the five phrases matching the response emotion. We used phrases for the four emotions joy, shame, sadness, and anger from our generated phrases. These emotions were chosen to best match the outcomes in 4.3 using response specified in [82]. Five phrases for each emotion were chosen to control for potential latent features fromo repeating the one melody, and to add variety to the robot's response in an effort to prevent tiring the user with the same sounds. We selected 5 of the 15 potential phrases for each trial by limiting length to be between 4 and 10 seconds. This restricted the variance to be less than 2, requiring the weighted mean emotion rating to fall within the correct quadrant of the wheel. The selected phrases were the ones with the smallest difference between the actual emotion and mean rated emotion.

In the notes group, the gestural response was accompanied by playing one musical note. Each emotion was randomly assigned one pitch from the midi pitches 62, 65, 69, and 72. The notes were chosen as they are in both the male and female vocal range and a similar pitch range to the EMP. We chose not to alternate timbre (the audio features outside pitch) for this group, as the EMP group already contained significant timbre variety. This assignment remained consistent throughout the experiment to maintain a relation between the sounds and the outcome. For each pitch, five different audio files were available to be selected, each with a different instrument timbre and length (varying from 2-5 seconds), to provide variety similar to that of the five different EMP phrases available for each emotion. Finally, in the gesture group, the gesture was performed in silence.

We created a gesture for each of the emotions joy, shame, sadness, and anger. The gestures were designed according to the table of emotion-specific nonverbal behaviors provided in [82] as well as our own post-hoc overview of discriminative body movements and poses. This approach has been used before in designing emotional robot gestures [159].



Joy

Shame

Sadness

Anger

Figure 4.1: Arm example poses passed through during emotional gestures

For the humanoid embodiment, we were able to incorporate more specific body language such as forming hands into fists and simulating crying.

For the Joy gesture the arm is lifted up high, making three quick upwards movements alternating which side it faces. The humanoid lifts both of its arms up and waves them back and forth, repeats this motion with its arms higher, and finally jumps into the air. For Shame the arm slowly bends down and away from the camera to one side, while the humanoid looks to one side and moves its hand to cover its face. For Sadness , the arm slowly bends down while still centered with respect to the camera, while the humanoid falls to its knees and covers its face with both hands. The Anger gesture has the arm first lean downwards and make two fast lateral movements, and then lean upwards to make two more fast lateral movements. The humanoid raises its fists into the air and push its torso forward. Examples of poses encountered during each gesture are shown in Figure 4.1 and 4.2.

Interaction Design

Our experiment required participants to perform a pattern learning and prediction task collaboratively with a robot. This is followed by two commonly used surveys; Schaefer's

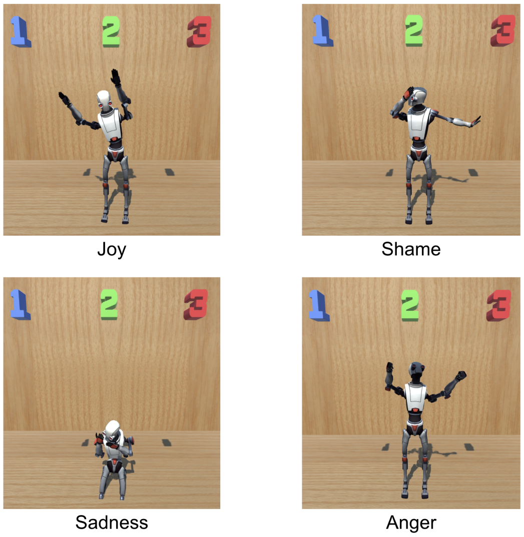


Figure 4.2: Humanoid example poses passed through during emotional gestures

survey for robotic trust [168], and the Godspeed measurement for Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and the Perceived Safety of Robots[63].

The study process followed 5 steps for each participant:

1. Consent form and introduction to online form
2. Description of the pattern recognition task
3. 20 Trial Pattern Recognition Tasks
4. 80 Pattern Recognition Tasks, recorded for data
5. Godspeed and Schaefer Trust Survey (order randomized per participant)
6. General comments and demographic information

The pattern learning method was originally created by Dongen et al. to understand the reliance on decisions and develop a framework for testing different agents[169]. Since then it has been re-purposed many times, including for comparing the dichotomy of human-human and human-automation trust [170], as well as the use of audio by cognitive agents[171].

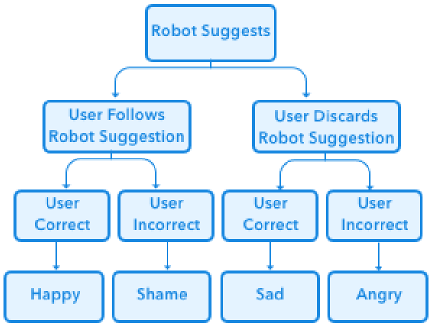


Figure 4.3: Robot Arm Emotional Response

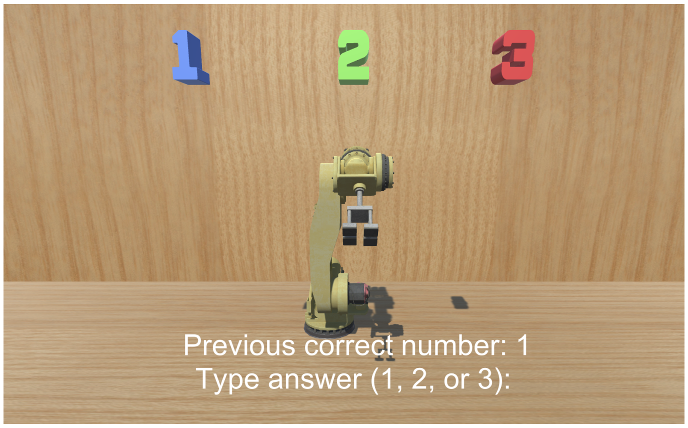
After collecting the consent form, participants went through a description of the task, followed by 20 trial questions to teach them the process. This was followed by the recorded and analyzed 80 questions. We finally allowed participants to add any general comments about the study or robot.

We modified the original pattern recognition task, asking participants to correctly predict the next number in a sequence advised by an animated model of a robot on a computer screen. Participants were told beforehand that humans and the pattern recognition software tend to be about 70% accurate on average, which has been shown to cause humans to alternate between relying on themselves and a decision agent. No further information was provided to the participants about the sequence's structure. The sequence was made up of a repeated sub-sequence that was 5 numbers long, containing only 1, 2, or 3 (such as 3, 1, 1, 2, 3). To prevent participants from quickly identify the pattern, 10% of the numbers in the sequence were randomly altered. Participants first completed a training exercise to learn the interface, in which a sub-sequence was repeated 4 times (20 total numbers). Then participants were informed that a new sequence had been generated for the final task. This was generated in the same way, using a new sub-sequence with 16 repetitions (80 total numbers). Before the user chose which number they believed came next in the sequence, the robot would suggest an answer, with the robot being correct 70% of the time. This process mirrors the process from the original paper [169].

Participants interacted with a virtual 3-D model of the robot in an application designed in Unity [1]. This allowed us to have interactions and responses vary based on user choices, while leveraging the quantity of participants available for an online study. Each time a participant was asked to answer a question, the robot acted as a decision agent, pointing to an answer that may be correct or incorrect. The user would then type their answer using their computer keyboard.

The previous timestep's correct answer was displayed for participants at decision time to help them better keep track of the pattern during the animated robotic movements. We required participants to submit their answer after the robot finished pointing to its prediction, which took between 2.5 and 4.5 seconds. This also forced participants to spend time considering their decision given the robot's recommendation. The robot would respond to the user's choice depending on the outcome and the condition of the experiment, as shown in Figure 4.3.

An example image of the robotic arm interface is shown in Figure 4.4.



Previous correct number: 1
Type answer (1, 2, or 3):

Figure 4.4: Example image from the robot interaction application

Participants

For each of the studies, we recruited 46 participants through the online survey platform Prolific[2] for a total of 92 participants. The participants ages ranged from 19 to 49, while

[1] https://unity.com

[2] https://www.prolific.co/

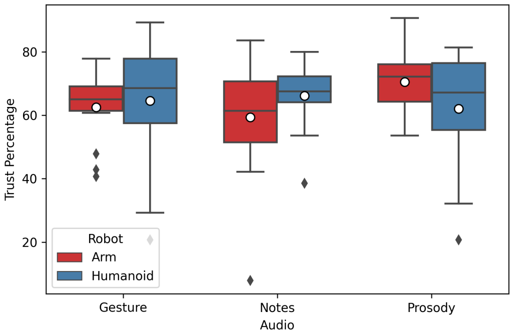


Figure 4.5: Box plot of Trust metrics. White dot indicates mean, middle line is median, and black dots are outliers.

the mean age was 25, with a standard deviation of 7. Participants were randomly sorted into one of the categories - EMP (15 participants), single-pitch audio (16 participants), and no audio (15 participants). Each experiment took approximately 30 minutes to complete. Participants were paid $4.75USD.

4.1.2 Results

To answer our research questions we used metrics from the trust survey, Godspeed measure, and the amount of times participants accepted the robot's suggestion. Research question 1 analyzes the results from the trust survey, while research question 2 focuses on the Godspeed metrics. Research question 3 compares the results from the trust survey with participant choices throughout the experiment.

RQ1: Trust Recovery for Industrial Arm

We first calculated Cronbach's alpha for each metric in the trust survey, which gave a high reliability of 0.92. We then calculated the overall trust score by inverting the negatively phrased questions and then generating the mean for each individual participant, resulting in a final trust percenttage. The mean trust of each group was EMP 71%, notes 57% and gesture 62% (see Figure 4.5). After running a one-way ANOVA the p-value was significant, p=0.041. Pair-wise t-tests between groups' trust rating gave the results: notes-gestures p=

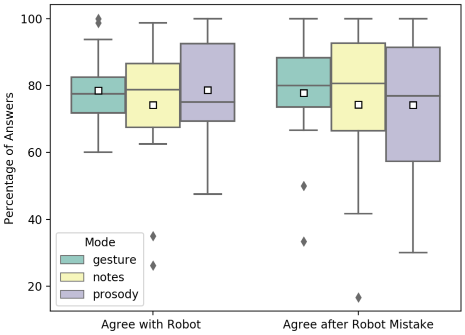


Figure 4.6: Robot Arm Box plots showing percentage of answers agreeing with the robot overall and after the robot made a mistake (means indicated by white squares)

0.46, notes-EMP *p*=0.025, and gesture-EMP *p*=0.025. This supports our hypothesis that trust would be higher from the arm using EMP.

We also evaluated trust based on participants' actual use of the system. The percentage of answers for which users agreed with the robot for each group are plotted in Figure 4.6. We performed a one-way ANOVA test to test whether there was a significant difference in this metric between groups, *p*=0.68, which was not significant.

To compare trust recovery after mistakes between groups, we analyzed the percentage of times each user agreed with the robot immediately after an instance of following the robot's incorrect suggestion. The results are plotted in Figure 4.6. The one-way ANOVA test yielded *p*=0.87, which was not significant.

RQ1: Trust Recovery for Humanoid

Cronbach's alpha for the humanoid trust survey was 0.89, showing a high internal consistency. We followed the same procedure to calculate the trust scores, with the means 0.63 for notes, 0.64 for gesture and 0.66 for EMP (see Figure 4.5). Running a one-way ANOVA and pair-wise t-tests showed no significance *($p > 0.05$)*.

Figure 4.7 shows the results for percent agreement with the robot, and percent agreement with the robot after it made a mistake. A one-way ANOVA between groups for

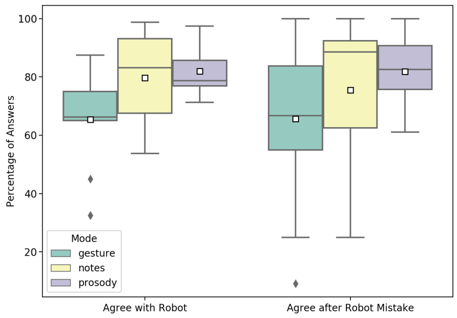


Figure 4.7: Humanoid box plots showing percentage of answers agreeing with the robot overall and after the robot made a mistake (means indicated by white squares)

percentage of answers in which users agreed with the robot yielded *p*=0.0039, which was significant. A 2-tailed t-test between each pair of groups had significant results for gestures versus EMP at *p*=0.0021 and gestures versus notes at *p*=0.018. The one-way ANOVA for percent agreement after the robot's mistake was not significant, with *p*=0.13. We note that we did not remove outliers for these statistical tests due to the number of participants in each group.

RQ2: Anthropomorphism, Safety, Intelligence and Likeability

Research question 2 identified how EMP, notes or gesture alone varied each Godspeed metric for the arm and humanoid robot. Cronbach's alpha for the robotic arm result in Anthropomorphism (0.85), Intelligence (0.89) and Likeability (0.92), and all showed high reliability values above 0.85. Safety's coefficient was slightly lower at 0.75. For the humanoid calculating Cronbach's Alpha for anthropomorphism, intelligence and likeability gave 0.80, 0.90 and 0.88 respectively, demonstrating high reliability. Safety's Cronbach alpha however resulted in 0.50 indicating the survey did not present internal validity. Due to the low internal reliability we chose not to analyze the safety results. This is discussed further in Section 4.1.3.

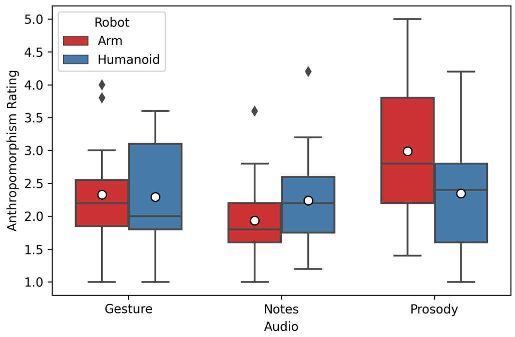


Figure 4.8: Box plot of Anthropomorphism, comparing humanoid and arm across audio types. White dot indicates mean, middle line is median, and black dots are outliers.

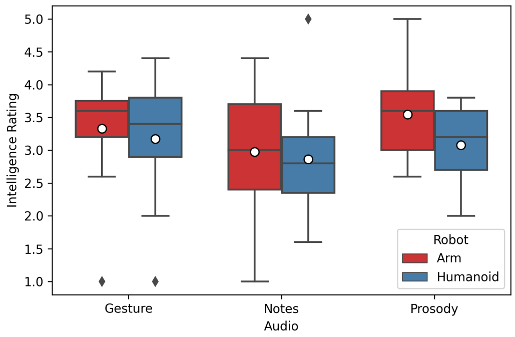


Figure 4.9: Box plot of Perceived Intelligence, comparing humanoid and arm across audio types. White dot indicates mean, middle line is median, and black dots are outliers.

RQ2: Safety, Intelligence and Likeability for Industrial Arm

We first performed a one way ANOVA for each category, which showed no significant results. Performing paired t-tests with Holm–Bonferroni corrections showed significance for anthropomorphism between EMP and gesture *(p = 0.048)* and EMP and notes *(p = 0.003)*. Likeability was also significant between notes and EMP *(p=0.048)*. Figures 4.8,4.9 and 4.10 show box plots for anthropomorphism, intelligence and likeability. This did not support our hypothesis as we were unable to show difference between audio types for safety or likeability across all categories.

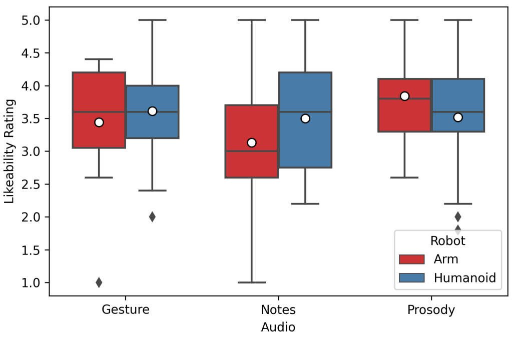


Figure 4.10: Box plot of Likeability, comparing humanoid and arm across audio types. White dot indicates mean, middle line is median, and black dots are outliers.

RQ2: Safety, Intelligence and Likeability for Humanoid

Across each audio category the humanoid achieved very similar results between the audio and gesture variables, with no significant difference. For example, likeability received ratings of 3.5, 3.52 and 3.61 for notes, gesture and EMP. These results indicated that the audio used made no difference to the perception of the robot. Figures 4.8,4.9 and 4.10 show the results for each metric.

RQ3: Trust Survey and Participant Choices

Research question 3 explored the relationship between the trust survey and participants actual choices throughout the experiment. We calculated the Pearson correlation coefficient between the final trust scores for the robotic arm, and the percentage of answers users agreed with the robot. The result was *r*=0.12, which indicates a weak correlation between the two metrics.

Arm - Qualitative User Comments

The free input textual comments provided by participants indicate that it was possible, in all groups, to perceive the emotions the robot was trying to convey. In the EMP group, one user said, ‘The arm seems quite emotional! When it’s right it is quite happy, but when it is wrong it gets particularly sad.’ In the notes group, a user said ‘When we got the right answer the robot seemed cheerful, as opposed to when we selected the wrong answer (based on the robot's recommendation) it seemed as if he was sorry for giving wrong suggestions. If I chose an option different than the robot's suggestion and its answer was correct, it seemed as if he gave the look of I told you the right answer!' And in the gesture group, one comment was 'the emotions were very easily perceivable.' Two participants in the notes group had negative comments on the audio response, describing it as 'horrible' and 'annoying', while one participant in the EMP group said the 'humming was annoying.' Several participants mentioned that the robot moved too slowly. Some comments mentioned having a hard time detecting any pattern in the sequence, while in others users discussed their strategies.

Humanoid - Qualitative User Comments

In the EMP group, one user said, 'It was clearly a robot(the cartoon), but the audio queues made it seem more humanlike,' with another user describing the robot as 'friendly.' However, another user in this group described the robot as 'irritating,' and another explained that it was pleasant at first but became annoying over time. In the notes group, two users used the phrase 'over the top' when describing the robot's reactions. Two other users mentioned that the robot seemed excited or like it was having a good time. One user said 'I feel like the sound effects aren't really necessary.' In the gestures group, one user said 'the robot seemed really happy when i got things right, but when i kept failing consistently i felt i was embarrassing it/letting it down, which added more pressure to me to getit [sic] right.' Two other users described a similar interpretation of the reactions. Two different users in this group mentioned that the robot seemed rigid or mechanic. Users' discussions of their strategies varied from trusting the robot most of the time, to trusting their own instincts more than the robot.

4.1.3 Discussion

Platform Specific Audio Design

Our goal for embedding EMP in robots was to develop and evaluate a non-language based form of audio communication, which could help avoid the uncanny valley. While we were successful in improving trust through embedding EMP in robotic arms, in humanoid robots we found no significant improvement in any category. This could be generally interpreted as meaning that audio does not alter humanoid perception as much as a lower degree of freedom, non anthropomorphic, robotic arms. It can also be claimed that our particular audio synthesis implementation did not lead to the desired results in humanoid robots, but that other future implementations might. The category humanoid robot is also very broad, with potential that the humanoid model we used was not able to be modified with audio, or that a feature such as the eyes dominated users perception. In any case, these results reiterated that audio must be carefully considered for every platform, without only reusing existing speech systems.

Godspeed

Comparing the Godspeed metrics, it was unsurprising to find that the addition of human vocalizations increased the Anthropomorphism of the arm. We had expected likeability to become higher, and while it was not a significant result, it would still be worth investigating further with more subjects. The most surprising result was that the pitch audio fell well below the median of gestures-only in every category. This may indicate that while EMP can lead positive outcomes, audio when implemented ineffectively has the capability to drastically reduce HRI metrics. The reason for this is likely due to the fact that the notes' sound was not related to the emotion being displayed by the gesture beyond remaining consistent throughout the experiment.

Measuring Trust

Users' ratings of trust in the survey did not strongly correlate with their actual behavior during the task in terms of how often they agreed with the robot's suggestions. This is consistent with the fact that while users reported significantly higher trust for audio with EMP, no significant differences were found in their actual choices during the interactions. A similar conflict between these types of metrics was found in the original decision framework paper [169], where higher reported trust in the arm did not always result in higher percent agreement with the arm.

We believe the primary reason for the contrasting rating for trust and how often participants agreed with the robot, is due to the multifaceted nature of trust itself. We contend that EMP is most impactful for changing ratings for affective trust, a type of trust that develops through emotion and social relationships. This contrasts with cognitive trust, which is based on a users actual willingness to trust or rely on a collaborator to perform a task [172]. In robotics, trust has similarly been broken into performance trust and moral trust [173]. Performance trust occurs where a human collaborator believes the systems is capable of performing the required action. The counter, moral trust, is a rating of the collaborators belief the robot desires to perform the morally correct task. While we make no claim that either measure we utilized to gauge trust directly correlates to a type of trust, we believe the trust survey is more likely to lead towards high ratings for affective or moral trust.

Measuring Perceived Safety

While the Godspeed survey has been extensively used in HRI with 1306 citations by December 2020, we believe an online study with animations may not have effectively used a perceived safety metrics. We found participants often described themselves as calm on the Godspeed scale, but also surprised, likely due to the online setting where surprising gestures did not change a participants self-perception as calm.

Limitations

This study was performed using virtual interactions with a robot, and 46 participants. It would be useful to investigate this further with a larger sample size, and to have participants interact with a physical robot for comparison. Additionally, more variations of robot responses could be compared and analyzed beyond the three that we investigated. For example, EMP of a human voice could be compared with that of musical instruments.

4.1.4 Conclusion

Our results support that when the robot arm model responded with EMP users reported higher trust metrics than when the robot responded with single-pitched notes or no audio. This supports our hypothesis that EMP has a positive effect on humans' trust of a robotic arm. We did not find significant results for likeability, anthropomorphism or perceived intelligence through EMP, although the arm with EMP did achieve higher means across both categories. In studies with a humanoid robot we found no significant changes in metrics, with audio seemingly have no impact on ratings. This indicates that audio design is a crucial step for human-robot interaction and can not simply be transferred between platforms without consideration of the broader impact.

4.2 EMP for Social Robotics

Text from this section has been published as:

Establishing Human-Robot Trust through Music-Driven Robotic Emotion Prosody and Gesture Richard Savery, Ryan Rose, Gil Weinberg, 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Dehli, India, 2019 [174]

Finding Shimi's Voice: Fostering Human-Robot Communication With Music And a NVIDIA Jetson TX2 , Richard Savery, Ryan Rose, Gil Weinberg, 17th Linux Audio Conference (LAC-19), CCRMA, Stanford University, USA, 2019 [151]

Our next study analyzing EMP in an embedded platform uses the social robot Shimi. Shimi moves with five degrees of freedom, and can play audio out of two speakers on either side of its head. Figure 4.11 shows an image of Shimi.

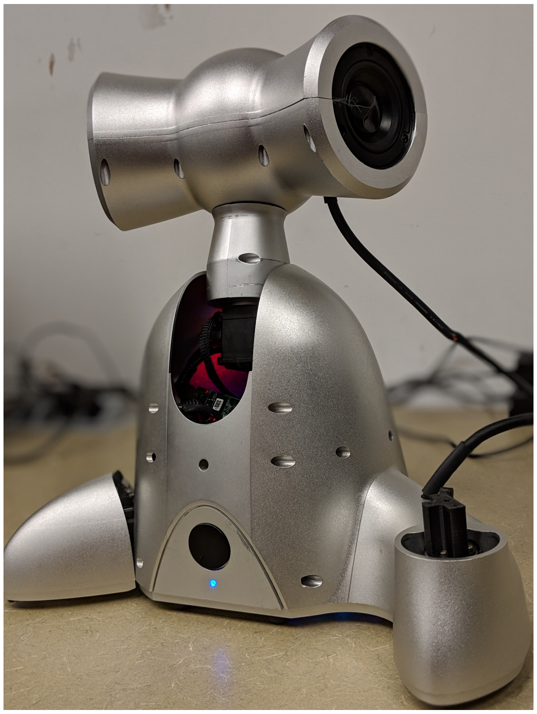


Figure 4.11: The social robot Shimi

4.2.1 System Overview

Shimi is an desktop musical robot companion, originally designed to act as an interactive music player. Prior work on Shimi focused on utilizing the sensors and computational power of a smartphone to explore the possibilities of personal robotics in a cost-effective way [175]. Other work on Shimi explored expressing emotion through gesture, informed by observations of human movement and emotion from Darwin [176, 177]. For the study conducted as part of this thesis we redesigned the internal components of Shimi by replacing the phone with a powerful microprocessor that can support embedded deep learning.

For many generative tasks, state-of-the-art performance depends on computationally heavy deep learning techniques. Embedded computing devices have only recently been developed with the GPUs necessary to perform complex deep learning inference in real-time. One such device is the NVIDIA Jetson TX2, an embedded system-on-module that runs Linux on a quad-core ARM processor, and features an 8GB GPU built on NVIDIA's Pascal architecture. This powerful and energy-efficient device greatly expands the capabilities of robots and other embedded applications alike through its ability to run both high CPU and GPU tasks, such as artificial neural networks, deep learning, and signal processing.

In this project, we embrace the non-human robotic identity of Shimi to explore methods of communication using Shimi's limited range of motion and music, in place of verbal language. This is realized through a voice generation system that utilizes deep learning to respond to human speech in an emotionally relevant manner, and a gesture generation system that uses both quantified emotion and Shimi's musical voice to craft robotic body language using Shimi's five degrees of freedom. This is combined with input analysis from a human respondent (see Figure 4.12.

Input Analysis

We programmed Shimi to analyze incoming audio streams using a combination of natural language processing (NLP) and raw audio analysis. Shimi features a Seeed Studio ReSpeaker Mic Array v2.0 [3], a four-microphone array with on-board processing that combines each microphone

stream and denoises the recording, emphasizing voice signals. No additional processing of input signals was added after the ReSpeaker processing, other than down-mixing to a single channel. Using the open-source hotword detection library `Snowboy`[4], Shimi responds to the phrase "Hey Shimi," and begins recording input audio. The Python phrase detection library `speech_recognition`[5] is then used to capture one phrase of raw audio.

[3] http://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

[4] https://snowboy.kitt.ai/

[5] https://github.com/Uberi/speech_recognition

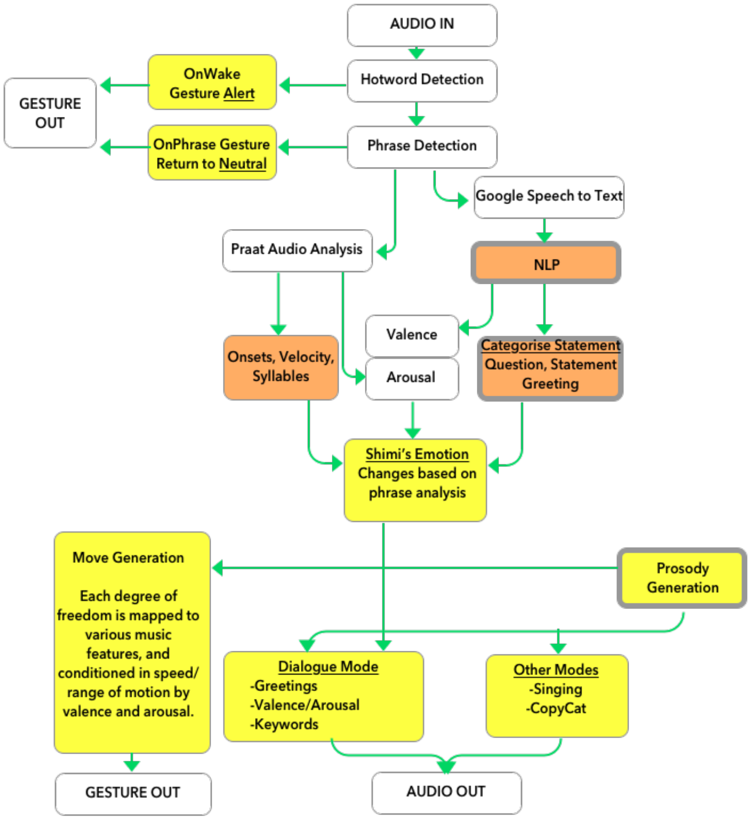

AUDIO IN
Hotword Detection
OnWake Gesture Alert
GESTURE OUT
OnPhrase Gesture Return to Neutral
Phrase Detection
Google Speech to Text
Praat Audio Analysis
NLP
Valence
Onsets, Velocity, Syllables
Arousal
Categorise Statement Question, Statement Greeting
Shimi's Emotion Changes based on phrase analysis
Move Generation
Each degree of freedom is mapped to various music features, and conditioned in speed/ range of motion by valence and arousal.
Prosody Generation
Dialogue Mode -Greetings -Valence/Arousal -Keywords
Other Modes -Singing -CopyCat
GESTURE OUT
AUDIO OUT

Figure 4.12: Shimi System Overview

Incoming audio is analyzed using the valence arousal model, whereby valence is the measure of the positivity or negativity of an emotion, and arousal is the measure of the energy of an emotion[178]. Raw audio analysis is used to find the arousal level, pitch, intensity and onsets. To do this we utilized `Parselmouth`[6], a Python library built on `Praat`[7]. We created custom metrics to analyze the input level based on analysis of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) data set [179]. RAVDESS includes 7356 audio files by 24 actors, each rated with an emotion independently validated by 10 participants. Our metrics were based on pitch contours and intensity levels found in the recordings. Figure 2 and 3 show analysis of the phrase *the dogs are sitting by the door* from the data set. Our metrics to measure arousal use the variety, level

[6] https://github.com/YannickJadoul/Parselmouth
[7] http://www.fon.hum.uva.nl/praat/

and standard deviation in intensity and the range, contour and standard deviation of pitch.

To measure valence we use the Natural Language Toolkit (`NLTK`) [180], a suite of Python modules for NLP. We calculate valence using a built in naïve bayes classifier trained on the `NLTK` data set of tagged phrases from social media. We also use the `NLTK` library for statement classification.

Shimi's Emotional Responses

Shimi maintains its own emotional state through each communication, tracked through a position in valence and arousal. Valence and arousal are both measured between -1 and 1. The current model gradually shifts the valence level towards that of the user while mirroring the arousal of the user. A negative valence statement from the user will cause Shimi to respond in a sad tone. Following positive statements from the user will gradually move Shimi towards positive responses. When starting Shimi begins with a valence of 0.5, equating to slightly happy.

Gestures

In human communication gestures are tightly coupled with speech [181]. Thus, Shimi's body language is implemented in the same way, derived from EMP and leveraging the musical encoding of emotion to express that emotion physically. The addition of gesture to EMP is crucial to add to the sense of embodiment, enhancing the sense that the audio is coupled to the robot itself. Music and movement are correlated, with research finding commonalities in features between both modes [182]. Additionally, humans demonstrate patterns in movement that is induced from music [183]. Particular music-induced movement features are also correlated to perceived emotion in music [184]. After a musical phrase is generated for Shimi's voice to sing, the MIDI representation of that phrase is provided as input to a gesture generation system. Musical features such as tempo, range, note contour, key, and rhythmic density are obtained from the MIDI through Python libraries `pretty_midi` [185] and `music21`[8]. These features are used to create mappings between Shimi's voice and movement: for example, pitch contour is used to govern Shimi's torso forward and backward movement. Other mappings include beat synchronization across multiple subdivisions of the beat in Shimi's foot, and note onset-based movements in Shimi's up-and-down neck movement.

After mapping musical features to low-level movements, Shimi's emotional state is used to condition the actuation of the movements. Continuous values for valence and arousal are used to influence the range, speed, and amount of motion Shimi exhibits. Some conditioning examples include limiting or expanding the range of motion according to the arousal value, and governing how smooth motor direction changes are through Shimi's current valence level. In some cases, the gestures generated for one degree of freedom are dependent on another degree of freedom. For example, when Shimi's torso leans forward, Shimi's attached head will be affected as well. As such, to control where Shimi is looking, any neck gestures need to know the position of the torso. To accommodate these inter-dependencies, when the gesture system is given input, each degree of freedom's movements are generated sequentially and in full, before being actuated together in time with Shimi's voice.

4.2.2 Experiment

We address two research questions, firstly, can we use EMP combined with gestures to accurately convey emotions? This was required to establish that our EMP generation system retains its ability to portray emotion when embodied in a social platform. We evaluate the effectiveness of the musical audio and generative gesture system for Shimi to convey emotion, specified by valence-arousal quadrants. Our second research question is whether emotional conveyance through EMP and gestures driven by music analysis can increase the level of trust in human-robot-interaction. For this we conduct a user study to evaluate EMP

[8] https://github.com/cuthbertLab/music21

and gestures created by our new model in comparison to a baseline text-to-speech system.

We designed an experiment to identify how well participants could recognize the emotions shown by our EMP and gestural emotion generator. This part of the experiment aimed to answer our first research question, can EMP combined with gestures accurately portray emotion. After watching a collection of stimuli, participants completed a survey measuring the trust rating from each participant. This part of the experiment was designed to answer the second question, can emotion driven, non-semantic audio generate trust in a robot.

We hypothesized that through EMP accompanied with low-DoF robotic gesture humans will be able to correctly classify Shimi's portrayed emotion as either happy, calm, sad, or angry, with an accuracy consistent with that of text-to-speech embodied in a robotic platform. Our second hypothesis was that we will see higher levels of trust from the Shimi using non-speech.

Stimuli

Table 4.2: Experiment Stimuli

Name	Audio	Stochastic	Experimental
Audio Only	X		
Stochastic Gesture, audio	X	X	
Stochastic Gesture, no audio		X	
Experimental Gesture, audio	X		X
Experimental Gesture, no audio		X	X

The experiment was designed as a between-subjects study, where one group would hear the audio with the Shimi voice, while the other would hear pre-rendered text-to-speech. Both groups saw the same gesture and answered the same prompts. The text-to-speech examples were synchronized in length and emotion to Shimi's voice. The stimuli for the Speech Audio experiment used CereProc's Meghan voice[9]. CereProc is a state of the art text to speech engine. The text spoken by Meghan was chosen from the EmoInt Dataset

[9] https://www.cereproc.com/

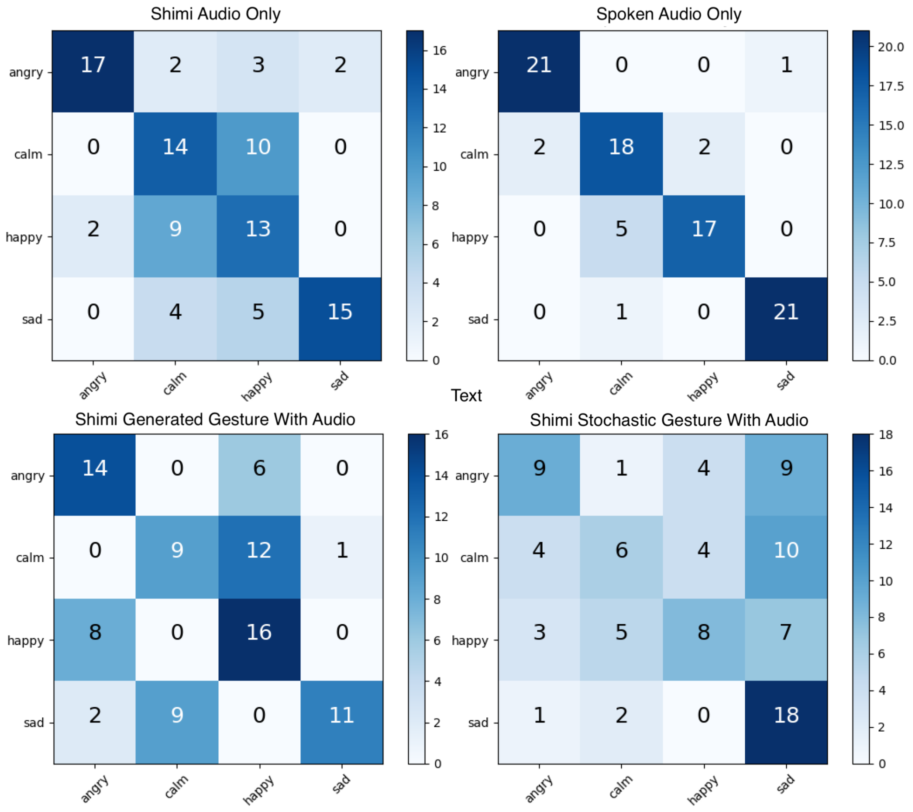


Figure 4.13: Confusion Matrix

[186], which is a collection of manually tagged tweets.

Emotion

The generated gestures were either deterministic gestures created using the previously described system, or deterministic stochastic gestures. Stochastic gestures were implemented by considering each DoF separately, restricting their ranges to those implemented in the generative system, and specifying random individual movement durations up to half of the length of the full gesture. The random number generator used in these gestures were seeded with an identifier unique to the stimuli such that they were deterministic between participants. Gesture stimuli were presented both with and without audio.

Procedure

Participants were gathered from the undergraduate student population at the Georgia Institute of Technology (N=24). Subjects participated independently, with the group alternating for each participant, culminating with 12 in each group. The session began with an introduction to the task of identifying the emotion displayed by Shimi. Participants responded through a web interface that controlled Shimi through the experiment and then allowed the user to select the emotion they thought Shimi was expressing. Stimuli were randomly ordered for each participant. Table 4.2 shows the order of stimuli used, each category contained 8 stimuli, 2 for each valence arousal quadrant. After identifying all stimuli participants were directed to a Qualtrics survey to gather their trust rating.

To measure trust, we used the Trust Perception Scale-HRI [55]. This scale uses 40 questions, each one using a rating scale between 0-100%, to give an average trust rating per participant. The questions take between 5-10 minutes to complete and include questions such as how often the robot will be reliable or pleasant. After completing the trust rating, participants had several open text boxes to discuss any observations in regards to emotion recognition, trust or the general experiment. This was the first time trust was mentioned in the experiment.

4.2.3 Results

Gestures and Emotion

After data was collected, two participant's emotion prediction data was found to be corrupted due to a problem with the testing interface, reducing the number of participants in this portion of the study to 22. First, we considered classification statistics for the isolated predictions of Shimi's voice and text-to-speech (TTS) voice. While TTS outperformed Shimi's voice (F1 score $TTS = 0.87$ vs. $Shimi = 0.63$), the confusion matrices show errant predictions in similar scenarios (see figure 4.13). For example, both audio classes

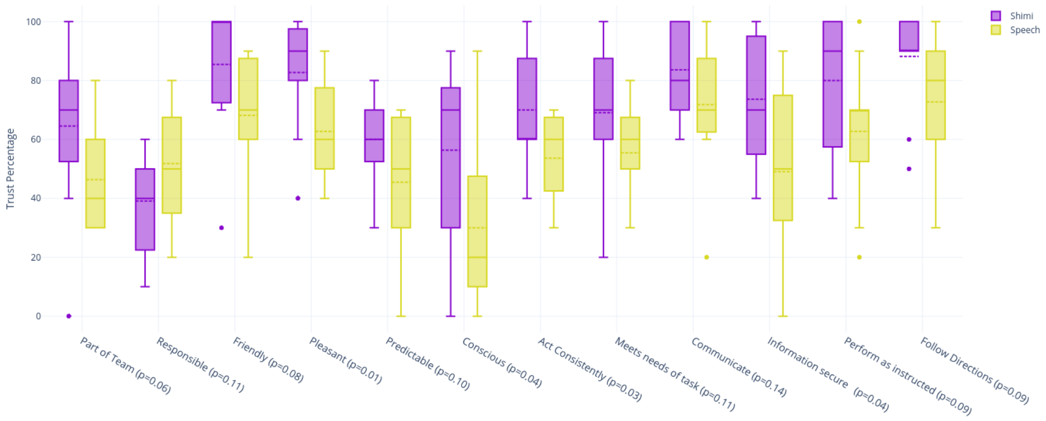


Figure 4.14: Questions with difference in mean over 10 %

struggle to disambiguate happy and calm emotions.

We hypothesize that adding gestures to accompany the audio would help to disambiguate emotions. To test that our gestures properly encoded emotion, we compared predictions for Shimi's voice accompanied by generated gestures with predictions accompanied by stochastic gestures, the results of which can also be seen in figure 4.13.

While the confusion matrices show a clear prediction improvement in using generated gestures over stochastic, the results are not statistically significant. A two-sided T-test provides a p-value of 0.089, which does not reject the null hypothesis at $\alpha = 0.05$. Disambiguities from the audio-only cases were not mitigated, but the confused emotions changed slightly, following other gesture and emotion studies [187].

Some experimental error may have accrued through the mixing of stimuli when presented to participants. Each stimuli was expected to be independent but some verbal user feedback expressed otherwise, such as: "the gestures with no audio seemed to be frequently followed by the same gesture with audio, and it was much easier to determine emotion with the presence of audio." The presentation of stimuli may have led participants to choose an emotion based on how we ordered stimuli, rather than their perceived emotion of Shimi.

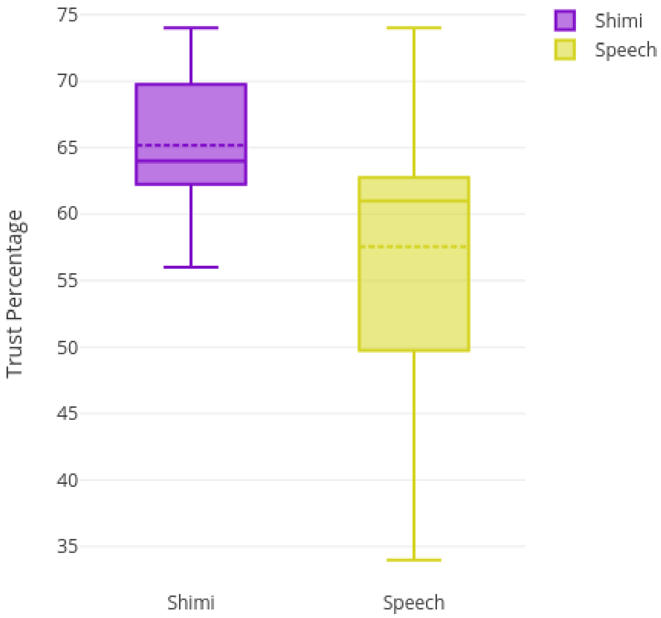


Figure 4.15: Participants Trust Mean

Trust

As per the trust scale, a mean percentage for trust was calculated on combined answers to 40 questions from each participant. A t-test was then run on each group mean. The average score variation between speech and Shimi audio showed a significant result (p=0.047), proving the hypothesis. Figure 4.15 shows the variation in average scores from all participants. The difference of mean between groups was 8%. Results from the text entries were positive for the EMP, and generally neutral or often blank for speech. A representative comment from the participants for the Shimi voice was "Seemed like a trustworthy friend that I would be fine confiding in."

Discussion

We were able to clearly demonstrate participant recognition of the expected emotion from Shimi, confirming our first hypothesis. Our model however did not perform completely as predicted, as audio without gesture lead to the clearest display of emotion. With a small

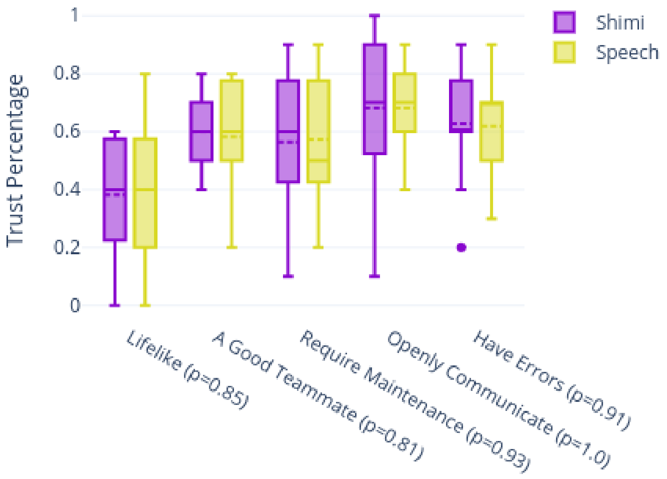


Figure 4.16: Not Significant Trust Results

sample size and a p-value close to being significant, we were encouraged by qualitative feedback that provided insight into the shortcomings of the gestures and gave us ideas for future improvements. For instance, emotions on the same side of the arousal axis were often hard to disambiguate. One participant noted that "it was generally difficult to distinguish happy and angry if there were no sounds (similar situation between sad and calm)", while another noted "I had some trouble discerning calm from sad here and there", and "without speaking, it was difficult to decipher between anger and excitement". The general intensity of the emotion was apparent, however." Certain movement features led to emotional connections for the participants, as demonstrated here: "generally, when Shimi put it's head down, I was inclined to say it looked sad. When it moved more violently, particularly by tapping [sic] it's foot, I was inclined to say it was angry or happy", "more forceful movements tended to suggest anger", and "When there was more severe motion, I associated that with anger. When the motion was slower I associated it with sad or calm. If the head was down more I associated it with sad. And I associated it with happy more when there was sound and more motion."

The trust perception scale is designed to give an overall rating, and independent questions should not necessarily be used to draw conclusions. However, there were several interesting results indicating further areas of research. Fig 4.14 shows all categories with a difference in median of over 10 %. Shimi's voice was crafted to be friendly and inviting and as expected received much higher results for pleasantness and friendliness. Unexpectedly, it also showed much higher ratings for its perception as being conscious. While further research is required to explore why this would occur, we believe that the question on consciousness of Shimi demonstrating a significant result shows that embodying a robot with a EMP (as opposed to human speech) creates a more believable agent. Figure 4.16 shows the categories with very similar distributions of scores. These include Lifelike, A Good Teammate, Have Errors, Require Maintenance and Openly Communicate. While further research is needed, this may imply that these features are not primarily associated with audio. Further work should be done to explore if the same impact can be found by adjusting audio features of a humanoid robot may also lead to interesting results.

In other future work we plan to develop experiments with a broader custom musical data-set across multiple robots. We intend to study emotional contagion and trust between larger groups of robots across distributed networks [188], aiming to understand collaboration and trust at a higher level between multiple robots.

Overall, our trust results were significant and showed that EMP and gesture can be used to generate higher levels of trust in human-robot interaction. Our belief that creating a believable agent that avoided uncanny valley was shown to be correct and was validated through participant comments, including the open text response: "Shimi seems very personable and expressive, which helps with trust".

4.3 EMP for Musical Robots Between Performance

Text from this section has been published as:

Before, Between, and After: Enriching Robot Communication Surrounding Collaborative Creative Activities, Richard Savery, Lisa Zahray, Gil Weinberg, Frontiers in Robotics and AI: Creativity and Robotics, 2021 [144]

In the previous sections we showed that EMP can have a positive effect on trust for robotic arms and social robotics. The rating for trust was based however on the overall perception of the robot, with no separation between features or exploration of how EMP alters individual tasks. In this section, we focus on how EMP can alter the perception of the actual task itself, such as the creative act of improvising music. In this way, we aim to identify if improved metrics such as trust, can correspond to improved ratings of the core functionality of a robot.

There is a growing body of work focusing on robots collaborating with humans on creative tasks such as art, language, and music [189, 190, 191]. The development of robotic functionalities leading to and following after collaborative creative tasks has received considerably less attention. These functionalities can address, for example, how a robot communicates and interacts with collaborators between musical improvisations, or before a piece begins or ends. Embodying a creative robot with speech capabilities that do not specifically address its creative capabilities risks distancing collaborators and misrepresenting artistic opportunities. In robotic literature this is referred to as the habitability gap, which addresses the problematic distance between a robot's implied capabilities and its actual potential output [1]. We propose that EMP could be particularly effective in human-robot collaboration in creative tasks, where emotional expression is at the core of the activity, and where subtle background conveyance of mood can enhance, rather than distract, from the creative activity.

We implement this system in a marimba playing robot, Shimon, and analyze the impact on users during creativity-based musical interactions. The musical tasks feature call and response musical improvisation over a pre-recorded playback. We compare the perception of common metrics of likeability and perceived intelligence, with the perceived creativity and preferences for interaction as well as Boden's creativity metrics [192]. We demonstrate that by using a creative communication method in addition to the core creative algorithms of a robotic system we are able to improve the interaction based on these metrics. Our implementation leads to the perception of higher levels of creativity in the robot, increased likeability, and improved perceived intelligence.

4.3.1 Experiment

For this experiment we embedded the EMP generation in our custom robotic platform Shimon. Shimon is a four-armed marimba playing robot that has been used for a wide range of musical tasks from improvisation [193] to film scores [194]. To visually show Shimon voicing the EMP we implemented a previous implementation of Shimon's gestures used for human language for hip hop [195].

For the experiment, we considered creativity using Boden's framework for computational creativity [192]. Boden considers creativity as a balance between novelty and coherence, with expressivity playing a significant role in the process. This concept draws on the notion that a new random idea could be considered novel but not creative, since it would lack coherence. Boden's framework was used to evaluate computational creativity in a number of previous works [196, 195].

We choose to compare EMP to a text-to-speech system for Shimon to further explore if EMP can lead to improved likeability and perceived intelligence. Speech is very commonly used in robotic collaborators [197, 198] and is likely the primary form of audio interaction. Speech is often described as a way for replicating human to human communication [12] and we believe would commonly be considered the default audio type for a robot such as Shimon. This contrasts with the robotic arm where speech is not commonly used, and instead gesture or simple audio warnings are more common.

Our experiment was designed to answer two research questions:

1 Can EMP improve the perception of a robot's creative output, as measured through novelty, coherence and expressivity when compared to a text-to-speech system?

2 Can EMP alter the perception of animacy, anthropomorphism, likeability and intelligence for a creative robot compared to a text-to-speech system?

To address these research questions we developed two exploratory hypothesis, extending the work of [1], where voices matching the mode of interaction will improve the interaction. For research question 1 we hypothesize that when communicating using EMP, Shimon will achieve higher ratings for novelty, and expressivity with a significant result, while coherence will not have significant difference. We hypothesize this will occur since EMP will increase the image of Shimon as creative agent, but not alter coherence. This aligns with our design goals of addressing the habitability gap and aiming for a robot that interacts in a manner that matches its performance. For research question 2 we hypothesize that there will be no difference in perception of animacy, and anthropomorphism, however EMP will achieve a significant result for higher likeability. We believe that the extra functionality implied by a text-to-speech system will enhance the perceived intelligence.

Experimental Design

We conducted the experiment as a between-group study, with one group watching robotic interactions with a text-to-speech system and the other with our generative EMP system. The study was set up as an online experiment with participants watching videos of a musician interacting with Shimon. For the text-to-speech we used Google API with a US female voice (en-US-Wavenet-E) [155]. We chose the voice model as it is easily implemented in real time and a widely used system.

The musical interactions involved six clips of a human improvising four measures, followed by Shimon responding with a four-measure-long improvisation. The improvisation was played over a groove at 83 beats per minute, resulting in the improvisation lasting for about 23 seconds. Each improvisation was followed by a seven-second gesture and response from Shimon, either using text-to-speech or EMP. Both the speech and EMP used three high valence-low arousal and three low valence-low arousal phrases. The EMP and text-to-speech was overdubbed after recording allowing us to use identical musical improvisations from the human and robot. For text-to-speech we used phrases that were designed by the author based on past interactions in rehearsal between human participants.

The high valence-low arousal text included the three phrases:

- Great work. What you played really inspired me to play differently. Could you hear how we were able to build off each others music?
- That was fun, it was good playing with you. I really liked hearing the music you played on keyboard, it worked well with what I played.
- Thanks so much for playing here with me, I thought what you played was really good. Let's keep playing together.

The low valence-low arousal text included the three phrases:

- Let's try it again soon, the more we play together the more we will improve. I'm going to listen to you really carefully next time
- That was a really good start, I enjoyed the way we interacted together. We should keep trying to work on it and get better.
- Did you listen to what I played? Do you think it worked well with what you played? The more we practice the better we can get.

Participants first completed a consent form outlining the process, and then read brief instructions on the experiment process. After watching three of the clips they were asked to rate them based on Boden's metrics, then repeated the process for the next 3 clips. Boden's metrics were rated on a seven point sliding scale. Participants were explicitly asked to rate the musical improvisation from the robot for each metric. Clips were randomly ordered for each participant. Additionally, a seventh clip was added as an attention check, which included an additional video. In this video, instead of sound, participants were asked to type a word that was asked for at the end of the survey.

After watching each interaction, participants rated animacy, anthropomorphism, likeability and perceived intelligence using the Godspeed measure [63]. Each metric contained four or five sub-questions, which were averaged to give an overall rating. To conclude the experiment, participants answered demographic questions and were given an open text response to comment on the robot or experiment.

We used Amazon Mechanical Turk (MTurk) to recruit participants who then completed the survey through Qualtrics. MTurk is a crowd-sourcing platform created by Amazon that allows individuals and businesses to hire users to complete surveys. Participants were paid $2.00 upon completion of the survey, which took around ten minutes. We allowed only MTurk Masters to participate, and required a successful task rate of 90%. We also monitored time to complete overall, and time spent to complete each question. We recruited 106 initial participants, four of whom failed the attention check. An additional two participants were disqualified as they completed the survey in under five minutes. As participants failed the attention check a new spot was immediately opened allowing us to reach 100 participants. In total we included data from 50 participants who heard the text-to-speech system and 50 who heard the EMP system. The mean age of participants was 44, ranging from 25 to 72, with a standard deviation of 11. The majority of participants were based in the United States (89) with the remaining in India (11). We found no difference in comparisons of the results between each country. Considering the gender of each participant, 39 identified as female, 60 as male and one as non-binary.

4.3.2 Results

Our analysis was conducted with a Jupyter Notebook, running directly on the exported CSV from qualtrics. Libraries for analysis included NumPy, and SciPy.stats.

Creativity

Our first research questions focuses on analyzing the creativity metrics, coherence, novelty, quality and expressivity. EMP had a higher mean for coherence 4.80 (*std = 1.31*), novelty 5.18 (*std = 1.30*), and quality 4.95, (*std = 1.68*) compared to speech with the means 4.19 (*std = 1.56*), 4.64 (*std = 1.24*), and 4.14 (*std = 1.37*). EMP had effect sizes of 0.40 for coherence, 0.43 for novelty, and 0.56 for quality indicating a medium size effect calculated using Cohen's D. For expressivity, EMP had an effect size of 0.25, indicating a small effect size. After conducting a pairwise t-test across categories were significant with the results, coherence (p = 0.041), novelty (p = 0.040), and quality (p = 0.014). After a Bonferroni-Holm correction for multiple comparisons, only quality remained significant with (p = 0.014) while coherence (p = 0.12) and novelty (p = 0.12) where no longer significant. For expressivity, EMP only had a slightly higher mean which was not significant ($p > 0.05$). Figure 4.17 shows a box plot of all Boden's metrics.

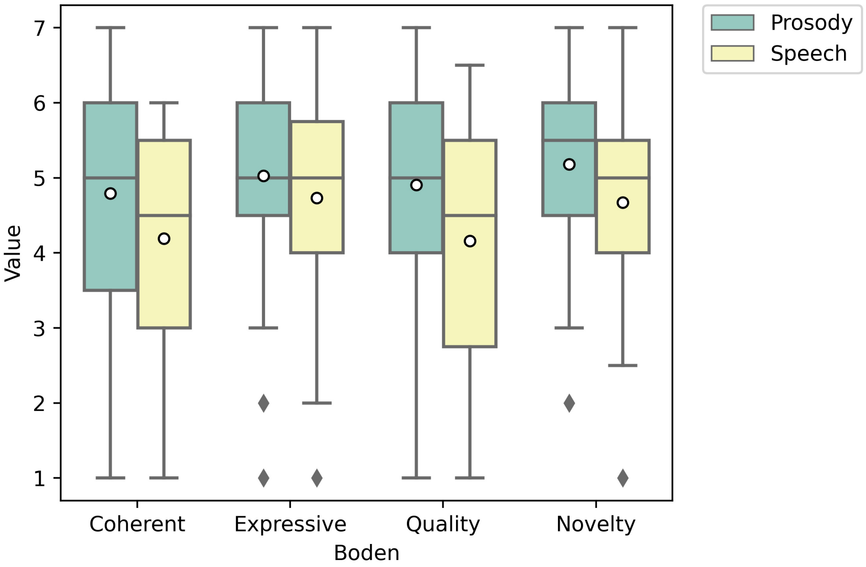


Figure 4.17: Box plot of Boden's Creativity Metrics. Quality is significant, p = 0.014.

Godspeed

For the Godspeed metrics we first calculated Cronbach's alpha for each question subset. This resulted in animacy (0.86), anthropomorphism (0.88), likeability (0.92), perceived intelligence (0.89). This shows high internal reliability across all metrics. EMP had an effect size for each metric as animacy (0.16) , anthropomorphism (0.08), likeability (0.85) and perceived intelligence (0.54), measured with Cohen's D.

EMP had a slightly higher mean for animacy 3.56 (*std = 0.88*) compared to speech 3.44 (*std = 0.75*). EMP also had a slightly higher rating for anthropomorphism 3.14 (*std = 0.99*), compared to speech 3.08 (*std = 0.885*). After running a pairwise t-test neither animacy or anthropomorphism were significant. EMP had a higher mean for likeability, 4.38 (*std = 0.89*) compared to 3.94 (*std = 0.52*) and showed a significant result (p = 0.002) in a pairwise t-test, which remained significant after a Bonferroni-Holm correction for multiple comparison (p = 0.011). For perceived intelligence, EMP 4.10 (*std = 0.82*) outperformed speech 3.72 (*std = 0.70*), with a significant result (p = 0.014) which remained significant after correction (p = 0.042). Figure 4.18 shows a box plot of all Godspeed metrics.

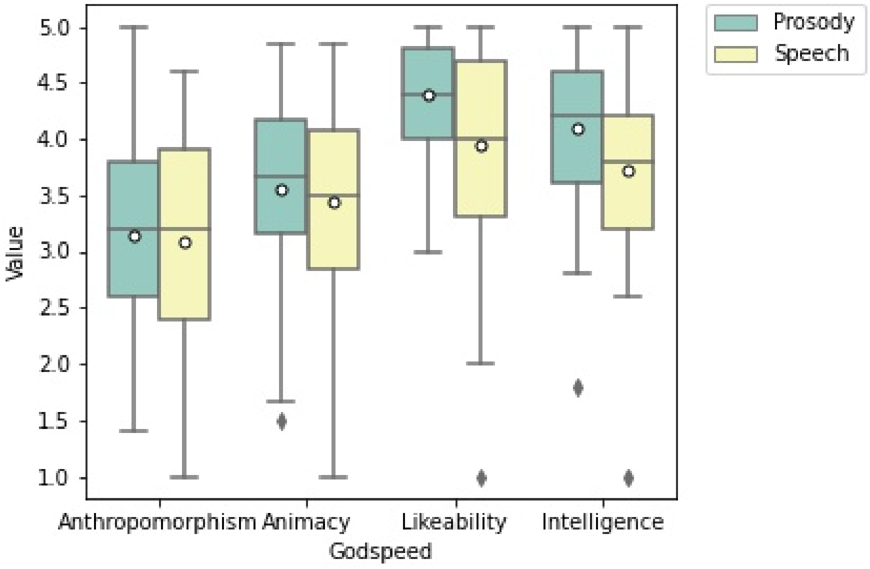


Figure 4.18: Box plot of Godspeed Metrics. Likeability and Intelligence are significant p = 0.011 and p = 0.042.

4.3.3 Discussion and Future Work

Research Question 1

Overall, our results indicated that the communication method outside of performance made a significant difference in participant ratings of creativity. The higher ratings for novelty and quality supported our hypothesis that EMP would outperform speech, however we did not expect coherence to improve with EMP as well. Surprisingly, we found no significant difference between voice type for expressivity and additionally expressivity only had a small effect size. This did not support our hypothesis as we had expected EMP to create the impression of a more expressive robot.

Further research is required to understand why the perception of expressivity, as a creativity trait, did not change based on the voice used. One possible reason is that participants believed a robot that could use language was capable of a wide range of expression, much like the addition of EMP. Alternatively, expressivity is a feature that is not easily altered by the form of interaction post-performance. Finally, it is possible that the movement or design of shown is inherently considered expressive with any type of audio added.

The relation between each creativity rating cannot be easily simplified, and there is no correct answer to what rating a performance should receive for coherence or novelty. We expected that the EMP system would receive higher ratings for novelty, but not coherence. We believe that the higher ratings for coherence may have come from the system acting as a unified robot, with its communication functioning in the same manner as its performance.

Research Question 2

Our results for likeability matched our hypothesis that EMP would outperform speech. Perceived intelligence ratings however did not support our hypothesis as we had predicted language would be interpreted as having a higher intelligence. It was reasonable to assume that with text-to-speech and the ability to speak a language, Shimon would have been perceived as more intelligent. We found that the system with EMP was considered more intelligent, despite not communicating linguistically. This can be explained by the assumptions that moving towards the habitability gap will create a disjointed perception of the robot. A possible conclusion was that participants understood there was not a deep knowledge of language, whereas musical phrases implies a deeper musical intelligence. The finding that perceived intelligence was not lower for EMP is very encouraging for further use cases of EMP. While it requires further research, the knowledge that EMP can raise key performance metrics, trust, and make a more believe agent, without the cost of hampering the perception of intelligence implies a range of future possibilities.

Text Responses

We found no distinct variation in text responses between the speech and EMP group. Overall 92 participants chose to respond, with responses ranging from one sentence to four sentences. From the speech group only one participant mentioned the voice, writing "I enjoyed the robot, especially when she spoke to the pianist" (gender added by participant). In the EMP responses four participants mentioned the voice, but only in passing, such as the voice was "cute". The vast majority of response rated the musical responses and generations, with the majority positive such as "I liked the robot and I like the robots music more than the humans", and "Nice to listen to". The negative comments tended to focus on the inability of robots in general to play music or be creative such as "It could play notes, but it lacked creativity".

Limitations

We compared one text-to-speech system with one EMP system on one robotic platform. In future work we aim to compare further audio systems, to expand understandings of why different metrics showed significant results. It is possible that varying the speech used would alter the final ratings. Nevertheless, we believe that the range of metrics that did prove significant show that this is an important first step in understanding how communication between core creative tasks can shape the perception of a robot.

We were only able to compare two forms of communication in a the constrained scenario consisting of directly after a musical interaction. To restrict our experiment to two groups we did not compare EMP to sections where the robot did not interact at all. We believe that by its nature a robot such as Shimon is always interacting and its presence can alter humans actions [199], leading us to believe that no movement or audio is its own form of interaction. In future research we intend to analyze the impact of EMP compared to no interaction in a longer performance.

This study was conducted online through video, which comes with benefits and drawbacks. As we were running online we were able to gather many more participants than would have been possible in person. Similar HRI studies have shown no difference in online replication of certain studies [200, 201], and we believe our method was constrained to a point that would be replicated in an in-person study. We did not include a manipulation check in our study, however our analysis of the text responses indicated that participants did not identify the independent variable between groups.

The range of participants included in the study also adds some limitations. Our primary goal was to understand how changes to a creative system would generalize across a broad population. We did not factor in concerns between cultural groups that may take place, such as between Japan and USA [202], however our study did not find any significant variation between origin country. Additionally, our ability to generalize is restricted by only collecting participants on MTurk, who it has been shown do not always represent standard population samples, such as in the case of participants health status [203]. Finally, our sample size of 106 participants was under the total that would be required to detect an effect size of 0.50 with 0.80 power at an alpha level of 0.05, which requires a sample size of 128.

4.3.4 Conclusion

This section explored how a robot's response outside of its key creative task - such as musical improvisation - alters the perception of the robot's creativity, animacy, anthropomorphism, perceived intelligence, and likeability. Our research question focused on how EMP compared to text-to-speech in a creative system for each of these HRI metrics. Through using EMP, we were able to increase user ratings for the key creativity ratings; novelty and coherence, while maintaining ratings for expressivity across each implementation. Our results also indicated that by communicating in a form that relates to the robot's core functionality, we can raise likeability and perceived intelligence, while not altering animacy or anthropomorphism.

The results in this section highlight the range of possibilities arising from the use of EMP. Building on broad findings from the robotic arm and Shimi, this study indicates that using EMP can alter not only broader perception metrics such as trust, but also the perception of core functionality used separately to EMP. Our results also present wide ranging implications and future concepts for the development of creative robots. The importance of design outside primary tasks should not only be considered for creative robots, but across HRI. These findings indicate that embodiment and external design choices alter not only the impression of a creative robot, but the impression of its primary functions.

4.4 Conclusion

Considering our original research question we found significant results for both the social robot and industrial robot for trust. For the industrial robot we found no significant results for the humanoid robot across any metrics, indicating either our approaches to audio were not effective, or that in the case of a humanoid robot audio is less easily able to change the perception of the robot. For the musical robot we were able to increase user ratings for the key creativity metrics; novelty and coherence, while maintaining ratings for expressivity across each implementation. Our results also indicated that by communicating in a form that relates to the robot's core functionality, we can raise likeability and perceived intelligence, while not altering animacy or anthropomorphism for musical robot. Importantly, we found that the role of EMP and the impact it has on a system can differ drastically between systems and implementations, highlighting the need for deep consideration of audio design.

After comparing platforms we believe EMP has the most potential benefit to robotic arms. While the benefits carried to social and musical robots, these robots have not been dispersed widely in the general public, with multiple examples of successful research not turning into commercial adoption [204]. Robotic arms are however already widely adopted, with Barclays estimates that the sales of such co-robot arms will grow to 700,000 units per year by 2025. We also believe that EMP is uniquely beneficial to robotic arms, with the habitability gap between types of interaction and audio through speech most likely to create uncanny experiences. For these reasons the following studies in this dissertation will focus experiments on potential implementations of EMP in robotic arms.

CHAPTER 5
EMP FOR PERSONALITY

Personality has been utilized in human robotic interaction research, such as in works that embed human personality in a robot to drive certain reactions and uses [18]. Another common approach is using human personality to understand robot perception, such as the overall impact of the uncanny valley [205]. While emotion is considered a critical feature of personality and is intertwined with the definition of personality itself [19], less research has been conducted addressing the interaction of personality, emotion, and robotics. In previous chapters we developed the generative system for EMP and analyzed its use in individual robots. While our experiment with the social robot Shimi involved emotion responses to a human, these were all programmed with basic copying rules. This chapter examines how emotion can be incorporated as a choice made by the robot to alter and improve interaction, through acting with personality traits and analyzing a human's personality traits.

In this chapter, we consider links between two of the Big Five personality types - Neuroticism and Extraversion, adaptive and maladaptive emotion regulation strategies, and robotics. The Big Five is the most common measure of personality in psychology [206, 207] and is considered cross-cultural [208] with each trait representing discrete areas of the human personality [209]. The personality traits in the Big Five, also known by the acronym OCEAN, are Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Here, we focus on Neuroticism and Extraversion, which have shown robust and consistent findings in regards to their role in emotion regulation for a human's personality [20]. These personality traits lead to emotion regulation strategies that involve the human process of exerting control over the intensity and type of emotion felt and how that emotion is displayed [210].

We contend that different robotic personalities can be projected primarily through audio using EMP. These personality types are grounded in the use of emotion regulation strategies with different levels of Neuroticism and Extraversion. We hypothesize that varying robotic personalities will receive different ratings by different human personalities. We propose that by using a consistent emotion regulation strategy from the Big Five framework, robots will achieve higher likeability and intelligence than a control group. These questions lead to the central research question of this chapter:

RQ 3: Does a person's personality alter their ratings of different emotional responses portrayed through robotic EMP?

Using personality to drive robotic emotion regulations can have multiple broader implications. Developing an understanding of personality and emotion can lead to the design of deeper interactions between humans and robots and inform the creation of a new framework for emotion driven interaction, leading to future improved understanding of trust. Emotion regulation provides the opportunity to drive new areas in human-robot interaction and develop new knowledge regarding the mechanisms that underlie affect based interaction.

For the study, we embedded custom emotional gestures and emotionally driven non-linguistic audio in an industrial robotic arm. The robotic gestures were based on human body language poses and were validated before use. The audio system was based on an EMP engine that has been shown effective for robotic arm interaction [160]. Avoiding speech and language has many advantages when it is not required for the interaction, such as reduced cognitive load [211] and improved trust [174].

In this chapter, we present two studies aiming to evaluate our platform for future use of emotion regulation and personality. The first study considers how human personality generates preferences for robot personality. The second study further examines emotion regulation strategies and preferences when hearing EMP, focused on human personalities.

5.1 Background

5.1.1 Personality and Robotics

There are a variety of frameworks for the analysis of human personality in psychology literature, with the most common categorizations classifying personality between three and seven traits [212]. In human robot interaction literature, the term personality is not always used consistently and often lacks an agreed upon framework [213]. It is relatively common for HRI researchers to describe robot personality based on distinctive responses to stimuli, without basing their work on any specific personality model [214, 215]. A few studies have shown the potential of embedding psychologically driven personality models in human robot interaction [216]. These include aligning human and robot actions based on human personality [217], predicting the acceptability of a robot in a teaching environment [218], and understanding the impact of personality on understanding robot intentionality [219].

Emotion modeling has been incorporated into some robotic personality models. For example, [220] use custom, subjective variations in emotional response to create nine unique personalities. [221] and [222] developed a robotic personality based on the Big Five, while using emotional responses based on possible relations between each class of the Big Five and emotion. However, these projects stay in conceptual level, do not have roots in psychology literature, and have not been tested with human users.

5.1.2 Emotion Regulation Strategies for Robotics

Emotion regulation is the process of modifying both an internal feeling of emotion and our external expression of an emotion [223]. There are three core features of emotion regulation that separate regulation from common approaches to emotion in robotics. The first is regulation relies on an intrinsic or extrinsic activation of a goal to modify emotions [224]. The second feature emphasizes mentally engaging with the cause of the emotion and changing one's internal reaction [225]. The third feature relies on varying the length and intensity of an emotional reaction [226].

There is limited work on emotion regulation in robotics, which often focuses on signaling [227]. Signaling implies that an internal emotion used by a robot must match the external emotion shown by a robot. Emotion regulation, however, which is a key element of emotion in humans and has direct links to personality, is hardly addressed in HRI research. A meta-analysis of emotion regulation and the Big Five found 32,656 papers including reference to regulation strategies linked to personality [228]. These findings are not always consistent however both Extraversion and Neuroticism had robust findings across the survey.

The literature overall indicates that emotion regulation strategies can be generated based on personality for agents with high Neuroticism and low Extraversion or low Neuroticism and high Extraversion. [21] in particular, describe contrasting response types for positive and negative emotion. High Neuroticism and low Extraversion (HighN-LowE) personalities are consistently more likely to respond to positive stimuli with reduced valence emotions, such as relief, whereas low Neuroticism and high Extraversion (LowN-HighE) are much more likely to respond directly with Joy or Happiness. For negative stimuli, HighN-LowE have a much higher likelihood to show disgust, fear, or guilt, while LowN-HighE are more likely to express sadness directly. In this chapter, we utilize these approaches to present a LowN-HighE robot and a HighN-LowE robot, each capable of responding with a different range of emotions to stimuli. This creates personality models that are able to respond to positive or negative stimuli, with varying response types, allowing a positive response to take multiple forms. This is wide expansion of previous chapters, and presents a method to actually implement emotions in autonomous robotic systems.

5.2 Personality Experiments

5.2.1 Stimulus

To display emotion regulation strategies we combined our EMP with a new robotic emotion gesture model created by Amit Rogel. The movements for each joint were created by hand to match our EMP. The gestures were designed by studying traditional human body language postures. Human gestures were broken down into their fundamental movements based on [82] and [229]. These motions were then mapped to various joints on the robot. Most of these mappings involved designing erect/collapsed positions for the robot as well as forward/backward leaning motions to create a linear profile of the robot that matched human gestures (shown in figure 5.1). Figure 5.2 shows the robots joint labels and motion.

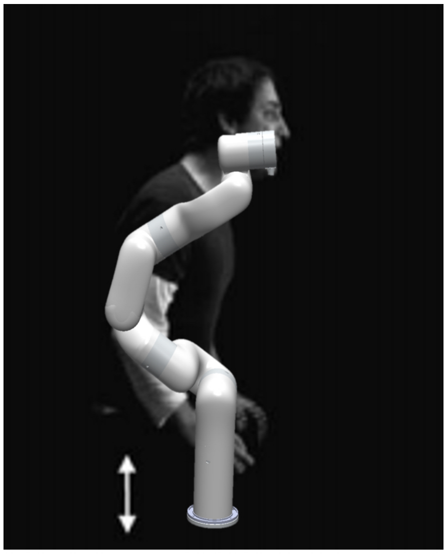


Figure 5.1: Example of robot creating a linear profile on top of [229]'s picture for fear

The robotic arm joints were primarily designed to mimic human motions. For example, the motion of joint 4 was designed to create erect and collapsed postures while joint 2 simulates forward and backward leaning motions. The position of fear as shown in figure 5.1, for example, depicts backward leaning of joint 2 leading to a slightly collapsed position of the full arm. To express sadness, the motions from [229], mapped joint 4 as a collapsed

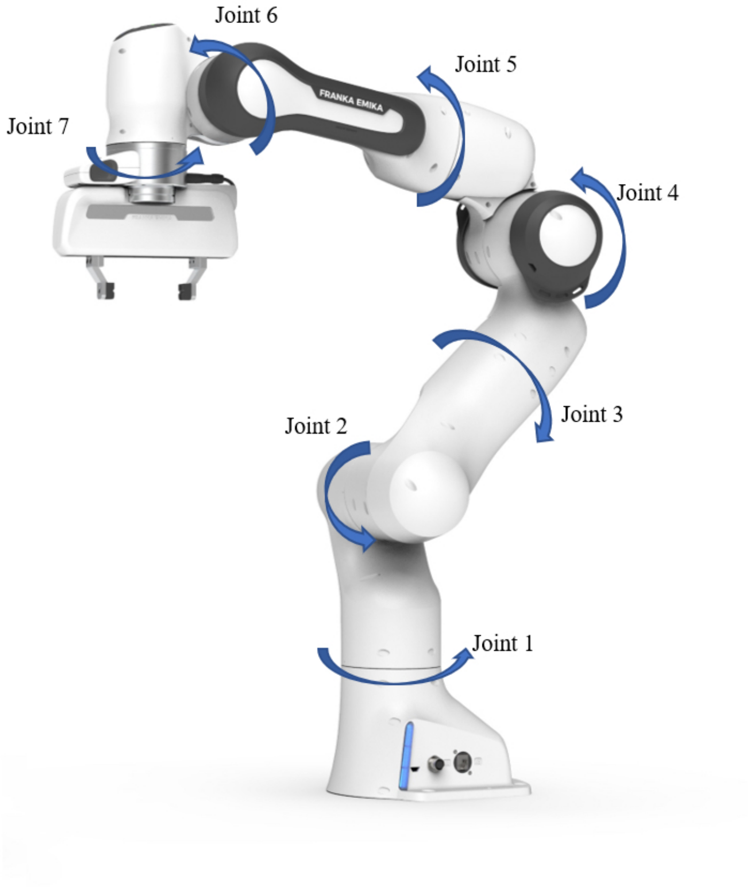


Figure 5.2: Franka Emika's Panda robot was used in study. Joint labels and movements are shown.

position while Joints 5 and 6 acted as head movements for the robot. To depict joy and admiration, joint 6 was designed to angle upwards to match the positive upright position of humans experiencing positive valance [82]. Joints 1 and 7 did not have specific emotion links, but were important in adding subtle changes to the robot to appear more human [1].

[1] The recordings of gestures used can be found here: www.richardsavery.com/personalitygestures

The gestures were hand designed while following the guidelines found in the table.

While human gestures informed the robotic arm's movement speed, rest times between movements and number of movements were designed to synchronize with the audio phrases to create a connection between EMP and the physical movements of the robot. After primary joint movements were established, smaller, subtle movements were added to some of the remaining joints to increase the animacy of the robot.

Validation

Human perception of the robotic gestures and sounds used in the experiments was validated in a user study. Each participant completed a survey containing 30 videos. Each video was approximately 8 seconds long and depicted a robot gesture and sound corresponding to a particular emotion. 17 different emotions were represented among the videos. After each video, participants were asked to identify the emotion they perceived, along with its intensity on a scale of 1-5, using the Geneva Emotion Wheel. One video was used as an attention check, which showed a robot gesture along with audio instructing the participant to select a particular choice. The validation used a total of 20 participants from Amazon Mechanical Turk. One participant was eliminated due to failing the attention check, leaving a total of 19 valid participants. Of these, there were 11 from the United States, 6 from India, 1 from Thailand, and 1 from Malaysia. 17 identified as male, and 2 as female. The mean age was 36.5.

We utilized two metrics to analyze the validity of the videos, based on [148] - the mean weighted angle of the emotions reported by participants and the respective weighted variance. Both of these metrics were weighted according to reported intensity, and were converted to units of emotions on the wheel. The average emotion error (absolute difference between weighted reported emotion and ground truth emotion) was 1.7 with a standard deviation of 1.1. The average variance was 2.8. All emotion errors were below 3.5 except for one video, which represented admiration and had an error of 5.0. Due to the overall higher emotion error and variance found for emotions in Quadrant 4 (positive valence, low arousal), we chose to keep this video in the study only for Experiment 2 and report results when it is included. However, we performed the statistical analysis with and without this video to ensure that it did not change the findings. These results show that participants were able to interpret the expressed emotions within a small range of error, making the videos suitable for use in the experiments.

5.2.2 Experiment One: Human and Robot Personality

The first experiment compares two robotic personalities driven by emotion regulation strategies, one with HighN-LowE, and the other with LowN-HighE.

Research Questions and Hypotheses

Research question 1 examines how the robot's personality alters its perception amongst all participants. This question does not consider the participants' personality type and instead aims to identify broad trends amongst all interactions. We will consider anthropomorphism, animacy, likeability, and perceived intelligence for each participant.

RQ 1) How does a robot's personality type as portrayed through emotion regulation strategies alter anthropomorphism, animacy, likeability, and perceived intelligence?

We hypothesize that the robot with LowN-HighE will achieve greater ratings for likeability and perceived intelligence, while we will see no difference in anthropomorphism and animacy across all participants combined. We believe that emotion regulation strategies matching LowN-HighE are conducive to immediate likeability in a short term experiment as they show less unpredictability. We believe predictability will also contribute to an increase in perceived intelligence.

Our second research question considers users in correlations between participants with LowN-HighE or HighN-LowE with robots with the same personality traits. Emotion regulation strategies are not as robustly found for humans with LowN-LowE or HighN-HighE, so we will not consider this category for this question.

RQ 2) How does a users' personality type impact their ratings of different emotion regulation strategies for anthropomorphism, animacy, likeability and perceived intelligence?

We hypothesize that each category will have a preference for the emotion regulation strategy that matches their own personality type for likeability and perceived intelligence, while there will be no difference for anthropomorphism and animacy. While the previous question described our belief that LowN-HighE would achieve better results, overall we believe that would occur largely to the addition of LowN-LowE or HighN-HighE, whereas each group individually will show significant variation in results.

Experiment Design

Participants first read a consent form and entered their names to confirm consent. They then completed the Ten Item Personality Measure (TIPI) [230], which gives the users personality with the Big Five emotion model. TIPI was chosen as it has shown strong convergence with widely used longer measures, and has been shown to effectively gather personality in online platforms such as Mturk [231].

The main section of the experiment involved participants seeing a photo followed by a robotic response. We used photos from the open effective standardized image set (OASIS) [232], which features a range of images tagged with valence and arousal ratings. We chose photos that clearly showed positive or negative sentiment, but also with a high standard deviation still within the bounds of positive or negative, implying a range of emotional response. We used a between experiment design, with participants randomly split into two groups, either seeing a robot responding to the stimuli with LowN-HighE or a robot responding with HighN-LowE. The responses were based on the response type described in Section 5.1.2, with each image returning an emotion based on the varying emotion regulation strategies. The same images were used for each robot personality type.

Figure 5.3 shows a sample sad image with a still of the robotic response. For each

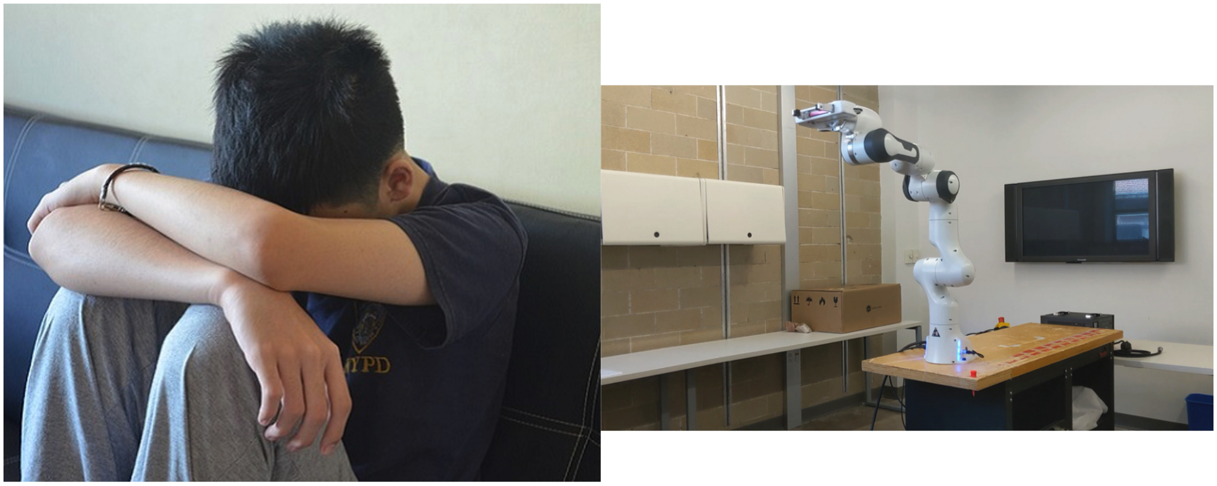


Figure 5.3: Sample Stimulus and Still of Robot Response

photo participants were asked to identify if the accompanying emotional reaction matched the image with a yes, no, or ‘other’ option. This was inserted to force participants to watch, as every expected response was yes. Stimuli were randomly ordered for each participant with an attention check also appearing randomly. The attention check involved a related image as well as audio requiring the participant to type a specific phrase in the selection box ‘other’.

Following reviewing the emotion stimuli participants were shown three text questions with an accompanying emotional response. The responses to each question were matched to expected responses by personality as found in work by [21].

1. How stressful was the task you just completed?
2. To what extent did you experience positive emotions?
3. To what extent did you experience negative emotions?

After viewing all stimuli, participants completed the Godspeed Questionnaire. Participants were asked to complete the survey while considering the robot across all videos shown for each image. Godspeed is a commonly used human-robot interaction standard for measuring anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots [63]. We chose not to ask participants about perceived safety as felt it was not relevant to the research question or reliably observed given the experiment design. The Godspeed Questionnaire involves 28 questions (22 without perceived safety), rating users' impression of a robot for terms such as Artificial to Lifelike, which combine to give the broader metrics. Following the Godspeed test, we collected participant demographic information including year of birth, country of origin, and gender. The combined study took no more than 15 minutes, with the average time to completion of 11 minutes. The survey form was hosted on Qualtrics.

We had 100 participants complete the study, of which 8 were eliminated due to failing an attention check, leaving a pool of 92. Of the 92 participants, the mean age was 42 with a standard deviation of 10 and a range of 22 to 69. 36 participants identified as female and 57 as male. Each participant was paid $2.00. 21 participants' country of origin was India, with the other 71 from the United States. We found no significant variation in responses from differences in countries of origin, gender or age.

Results

We first analyzed the participants' personality results and found the break down between Neuroticism and Extraversion as HighN-HighE n=11, LowN-LowE n=13, HighN-LowE n=27, and LowN-HighE n=36. For the Godspeed test, we first calculated Cronbach's Alpha for each category. The results for each category were: Animacy 0.83, Anthropomorphism 0.88, Likeability 0.92, and Intelligence 0.91. This indicates a high internal consistency across all survey items.

Research Question 1

The robot personality with LowN-HighE emotion responses had a higher mean for both likeability and perceived intelligence. After conducting pair-wise t-tests the results were significant for both categories; for likeability (*p=0.011*) and for perceived intelligence (*p=0.015*).

For likeability LowN-HighE had a mean of 4.191, a standard deviation of 0.684, and the confidence interval (3.903,4.480). LowN-HighE had a high effect size of 0.856. HighN-LowE had a mean of 3.606, a standard deviation of 0.924, and a confidence interval (3.272, 3.940). For the intelligence statistics LowN-HighE had a mean of 3.992, a standard deviation of 0.790, and the confidence interval (3.658,4.325). LowN-High had a high effect size of 0.741. For intelligence HighN-LowE had a mean of 3.406, a standard deviation of 0.919 and the confidence interval (3.074,3.737). For anthropomorphism and animacy the results were not significant ($p \geq 0.05$). These results proved our hypothesis and showed that the robotic personality type did alter the general populations' ratings for likeability and perceived intelligence. Figure 5.4 shows a box-plot of the results.

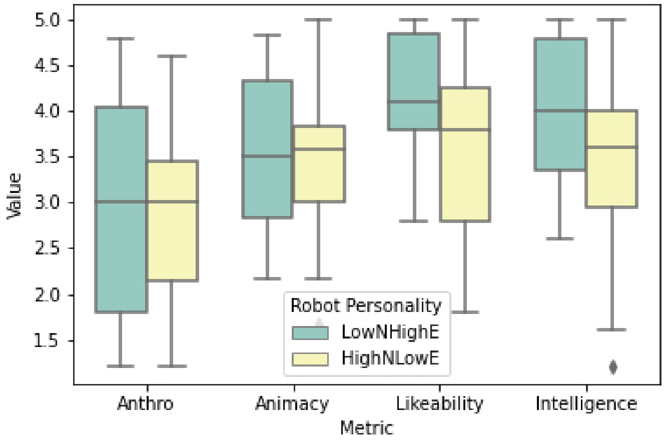


Figure 5.4: Comparing Robot Personality Across All Participants

Research Question 2

Both human personalities rated the robot with LowN-HighE higher for likeability, with a pair-wise t-test giving significant results for LowN-HighE (*p=0.025*) but not for HighN-LowE (*p=0.147*). Figure 6.4 shows an overview of these results. This partly supported the hypothesis with LowN-HighE preferring LowN-HighE, but without significant results for HighN-LowE. Likewise perceived intelligence rating was higher from both for LowN-HighE, but again only with significant results for LowN-HighE human personalities (*p=0.049*), and for HighN-LowE (*p=0.78*).

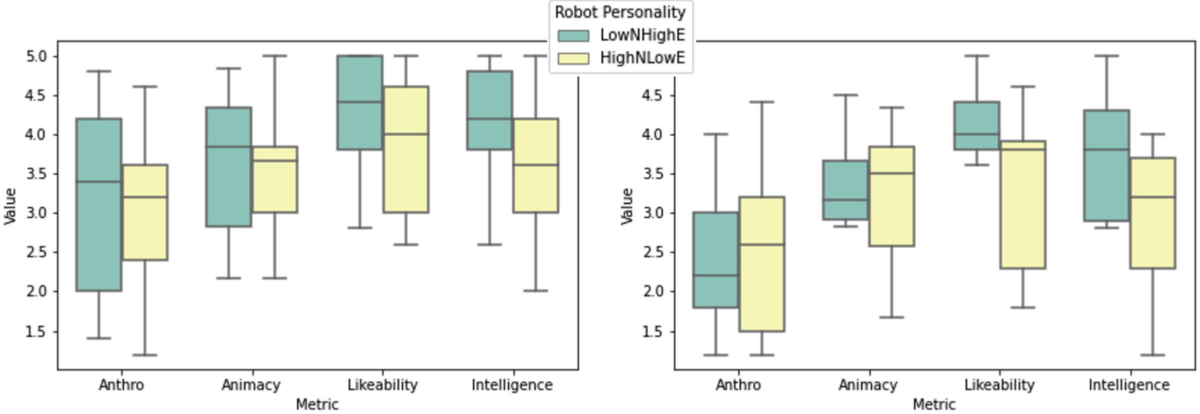


Figure 5.5: Comparing Human Personality Across Platform. Left indicates humans with LowN-HighE, right HighN-LowE

Contradicting our hypothesis, both animacy and anthropomorphism showed ratings for robot personality that matched that of the human personality. Users with LowN-HighE rated the robot with LowN-HighE better for both animacy and anthropomorphism although neither was significant (*p≥0.05*). HighN-LowE also rated animacy and anthropomorphism higher for the robot with HighN-LowE, with a significant result for anthropomorphism (*p=0.004*). Further discussion of these results is available in Section 5.2.4, including comparisons with the results from our second experiment.

Supplementary Results: Openness, Conscientious and Agreeableness

Our research questions focused on collecting and analyzing the personality traits Neuroticism and Extraversion, however standard personality measures for the Big-5 also include Openness, Conscientiousness and Agreeableness. Openness is linked to levels of curiosity and willingness to try new things; conscientiousness is considered a efficiency and organisation, while agreeableness is related to friendliness and compassion. As previously described these traits do not have consistent findings in relation to emotion regulation, nevertheless we believe analyzing their links to our other variables is worth consideration to guide future work.

Our results for Openness to experience matched expectations, with the more open a participant the more likely they were to rate both robot personalities as likeable and intelligent. Comparing openness and intelligence gave a Pearson's correlation coefficient of 0.4 with *p=0.002*, indicating a moderate positive relationship. Figure 5.6 shows the high and low openness trait for each metric.

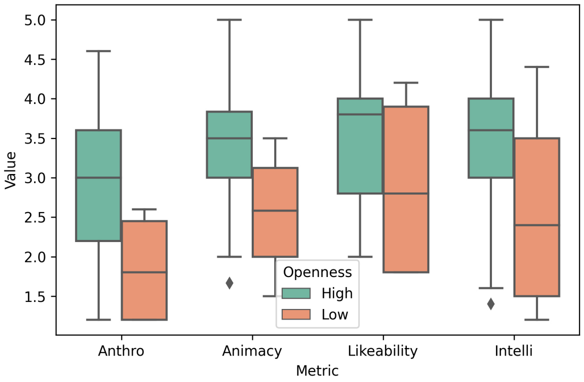


Figure 5.6: Openness to experience personality trait rating for each metric

While [231] found Mturk personality surveys gave accurate results, we believe TIPI was insufficient for measuring conscientiousness and could not draw any conclusions on the trait. TIPI includes two questions for measuring conscientiousness, asking for a self-rating of participants' dependability and carefulness. For Mturk we believe participants would be wary to mark either rating too low and risk their rating on the platform. This lead to a distribution with 88 participants rating themselves as highly conscientious and 5 giving themselves a low conscientious rating.

We found no relation between agreeableness and preferences for emotion regulation or robotic personalities. The Pearson correlation coefficient for each metric was: animacy

(0.136, *p=0.195*), anthropomorphism (0.46, *p=0.661*), likeability (0.195, *p=0.062*), and perceived intelligence (0.190, *p=0.069*).

This replicates common psychology findings, that find agreeableness plays a part in emotion regulation near exclusively in social emotion settings [233, 234, 235].

5.2.3 Experiment Two: Emotion Regulation Preferences

In the first experiment user's decide between two preexisting personalities in order to understand emotional regulation strategies. The second experiment aims to explore these strategies with participants creating a robot personality. Based on the user's desired reaction for the robots, the robot could be either high Neuroticism-low Extraversion (HighN-LowE), or low Neuroticism-high Extraversion (LowN-HighE). The second experiment further explores desired personality types in robots.

Research Questions and Hypothesis

The second experiment builds on our findings from the first experiment in hopes of finding a consistent regulation strategy. Now that participants can build a more ideal personality, we want to consider and further explore the most appropriate reactions.

Research Question 3) Do users consistently choose one emotion regulation strategy for all stimuli?

Research Question 4) How does a user's personality type impact their preference when comparing between two different emotion regulation strategies (HighN-LowE and LowN-HighE) for a robot?

We hypothesize that results will follow a similar trend to research question 2; each category will prefer an emotional response similar to their own personality type. Just like the second research question, LowN-Low-E and HighN-HighE were not included because they are not as reliable.

We also hypothesize that participants may prefer a more consistent robot personality, and therefore would tend to select videos from one of the two emotion regulation strategies.

Experiment Design

Like the first experiment, participants began the study by reading and signing a consent form and completing the TIPI. The experiment differs in the primary section, where participants were given a photo and two videos of the robot. The photos used were from the OASIS set of photos used in the first experiment. The photos include a variety of valance and arousal tags. For each image, two robot response videos were displayed. Each one of the videos was associated with an emotion, reacting to the stimulus as either highN-lowE or lowN-highE. Participants were asked which response they would prefer to see in the robot as a reaction to the given image.

Each participant was shown 14 photos with robot responses, as well as one additional photo that was used as an attention check. In this question, one video instructed the user to select the ”other” option and type a specific phrase. The ordering of the two video responses was randomized for each question, and the ordering of the questions was randomized for each participant.

We had 100 participants complete the study, of which 14 were eliminated due to either failing an attention check, or failing to follow instructions and selecting the ”other” option for a question that wasn’t the attention check. This left a remaining pool of 86. Of the 86 participants, the mean age was 41 with a standard deviation of 10 and a range of 25 to 68. 25 participants identified as female and 61 as male. Each participant was paid $1.50. This is less than Experiment One because this study was estimated to take less total time to complete. 23 participants’ country of origin was India, 60 from the United States, 1 from Libya, 1 from Thailand, and 1 from Singapore. We found no significant variation in responses from differences in countries of origin, gender or age.

Results

We first analyzed the participants' personality results and found the break down between Neuroticism and Extraversion as HighN-HighE n=4, LowN-LowE n=31, HighN-LowE n=15, and LowN-HighE n=36.

Research Question 3

To investigate whether participants tended to consistently choose one emotion regulation strategy over the other, we looked at the distribution for the difference between the number of times each participant selected from the two strategies. If participants consistently chose one of the two categories, we would expect to see a bimodal distribution, with some participants centered around more HighN-LowE choices, and others centered around more LowN-HighE choices. Figure 5.7 shows this distribution, plotting the number of HighN-LowE selections minus the number of LowN-HighE selections. It appears normally distributed, with a mean of -.70 and a standard deviation of 3.7. The confidence interval was (-1.48, .08). We performed a Chi-squared goodness of fit test between the observed data and a binomial distribution with probability 0.5. χ^2 was 4.3, which is less than the critical value of 6.6 (for 14 degrees of freedom). This supports that there is no significant difference between the observed distribution and a binomial distribution.

To determine whether there was a trend among all participants of one strategy over the other, we performed a 1-sample t-test on these differences, comparing against an expected mean of 0. The p-value was .083, which is not significant. The effect size was .19.

These results do not support that participants consistently chose one emotion regulation strategy over the other.

Research Question 4

We first performed a 2-tailed t-test for the percentage of times each participant selected the HighN-LowE robot response, comparing between the HighN-LowE participants and

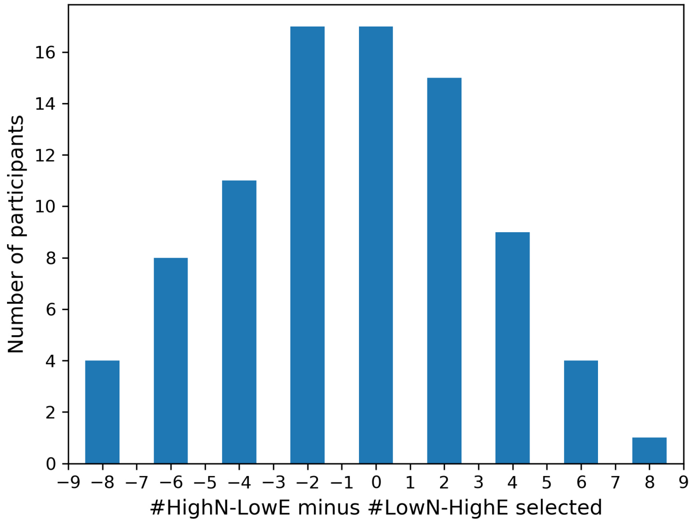


Figure 5.7: Distribution of participant video selections, difference between number of HighN-LowE vs. LowN-HighE selected

the LowN-HighE participants. For HighN-LowE participants, the mean was .45 and the standard deviation was .12, with confidence interval (.39,.51). For the LowN-HighE participants, the mean was .46 and the standard deviation was .14, with confidence interval (.42,.51). The result of the t-test was p=.81 which is not significant. The effect size was .07.

To investigate further, we performed two Pearson correlation tests. Both tested correlation with the percentage of times each participant selected the HighN-LowE robot response. We first tested correlation with each participant's neuroticism (N) score, with a result of r=.038, indicating no correlation. We then tested with each participant's extraversion (E) score, with a result of r=.049 also indicating no correlation. Figure 5.8 shows this relationship.

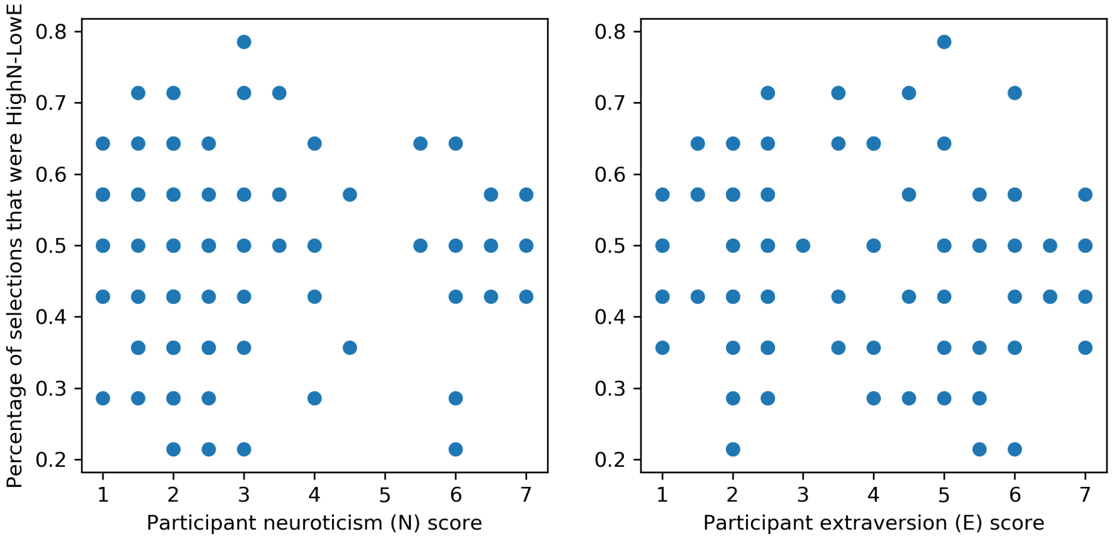


Figure 5.8: Scatter plots showing no correlation between participants' selection of emotion regulation strategy and Neuroticism/Extraversion

5.2.4 Discussion

Experiment One: Human and Robot Personality

We found LowN-HighE consistently more likeable for all users, with significant results for the LowN-HighE human with LowN-HighE robot. While we can not conclude why this is the case, we believe it may be due to the nature of short-term interaction. Especially in a single encounter, it is reasonable to assume that a robotic agent that shows higher extraversion and more emotional stability (through lower neuroticism) is more immediately likeable regardless of a user's personality.

LowN-HighE also received higher ratings for perceived intelligence across both personality classes. This indicates that perceived intelligence is much more than just the ability to accurately complete a task. All users almost unanimously rated the robot as correctly identifying the emotion, yet still found a significant difference in perceived intelligence. As for likeability, we believe this reduced intelligence rating is due to higher levels of emotional instability.

Contradicting our hypothesis, anthropomorphism and animacy ratings corresponded to human personality types, with HighN-LowE and LowN-HighE both rating their matching robotic personality higher. While we did not predict this, we believe this does make sense as users who see emotion regulation strategies closer to their own may be more likely to see anthropomorphic characteristics in a robot and more lifelike behavior.

LowN-LowE, HighN-HighE

Our core personality design involved HighN-LowE and LowN-HighE, however, we were also able to analyze LowN-LowE and HighN-HighE. Our sample size from experiment one was significantly smaller for both these groups (n=11 and n=13). Figure 5.9 shows the results for all personality types. LowN-LowE and HighN-HighE personalities are less common and less easily grounded in literature, so any conclusions from this data are not easily verified. However, there are some clear distinctions between comparisons of each human personality. HighN-HighE has almost no variation between robot personality with no significant results. This implies either that emotion regulation strategies do not impact this personality type, or that neither of our emotion regulation strategies strongly impacted HighN-HighE personalities. LowN-LowE personalities however did not have significant results for the LowN-HighE robot, for perceived intelligence (*p=0.48*) and likeability (*p=0.49*). This matches the results achieved for the general population and the LowN-HighE group. Despite these results, there is still future work required to draw any conclusions about LowN-LowE and HighN-HighE personalities and robotics.

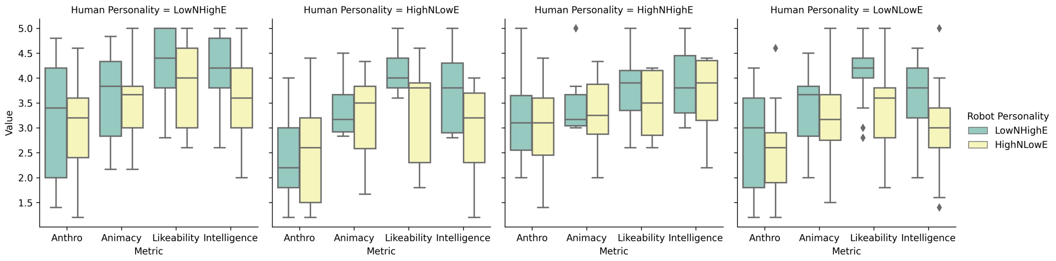


Figure 5.9: Comparing LowN-LowE and HighN-HighE

Experiment Two: Emotion Regulation Preferences

We did not find that participants consistently chose one emotion regulation strategy over the other, and did not find a correlation between their choices and their personalities. One reason for this could be that the first experiment looks at all the emotion regulations as a whole while the second experiment focuses on a user evaluating each strategy as a specific entity. The participants were making specific pairwise comparisons between videos, making the results fairly dependent on the specific gesture/audio sample pairs. This focus on a particular reaction may put more emphasis on a gesture and audio preference than a personality preference.

Additionally, only 14 comparisons were made in total. There were some high variances in the emotion validation, especially for emotions in Quadrant 4 of the Geneva Emotion Wheel. Even though the mean weighted reported emotions were generally close to their ground truth values, the high variance means that many video examples may be necessary in order for the results to successfully represent the emotion over specific gesture/audio samples. This may have caused the users to have trouble properly identifying the personalities.

Limitations

While attempting to control for all weaknesses in the study, there are several limitations that are worth describing. We did not collect information on participants on how they perceived the personality of each robot, so do not have a firm metric that the robot was believed to be a certain personality. This however was a considered decision; it has been repeatedly shown that untrained humans are inaccurate at predicting others' personality types through observation, especially over short interactions [236, 237]. Nevertheless, future work attempting to identify how emotion regulation in robotics portrays a personality type to users would be of benefit.

Our study used videos of the robots interacting instead of in person participation. We believe for this experiment this did not alter the end results and improved overall outcomes as we were able to recruit many more participants than would be possible in person. Multiple past papers have shown no significant variation in results when a participant is watching a robot on video compared to in person [200, 201]. In future work, we expect to apply lessons learned from these studies to in person experiments and interactions and believe lessons learned from video will apply to in person studies.

5.3 Conclusion

Considering our third research question, the findings suggest that all human personalities prefer to interact with robots showing low Neuroticism and high Extraversion over the short term. No significant results were found regarding the perception of anthropomorphism and animacy. The chapter overall presents a new framework for developing emotional regulation and personality strategies for human-robot interaction. It explores how the Big Five personality traits can inform future designs of emotion-driven gestures and sound for robots. In particular, it studies the interplay between human and robotic Neuroticism and Extraversion and their effect on human perception of robotic personality.

This chapter has demonstrated possible future directions for implementing autonomous EMP reactions into robotic systems. Whereas the previous chapter focused on confined, scripted interactions, this chapter has demonstrated how these interactions can be expanded and personalized for robotic arms. Key broader contributions include the development and implementation of novel affect and personality models for non-anthropomorphic robotic platforms. Other contributions include a groundwork understanding of emotion regulation strategies in human-robot interaction and novel insights regarding the underlying mechanism of emotion and affect in robotics.

CHAPTER 6
EMP FOR ROBOTIC GROUPS

Text from this section has been published as:

Emotion Musical Prosody for Robotic Groups and Entitativity, Richard Savery, Amit Rogel and Gil Weinberg, 30th IEEE International Conference on Robot & Human Interactive Communication, 2021 [238]

In the following chapter, we aim to explore how EMP can be expanded beyond dyadic human-robot interactions, exploring the role of EMP in groups of robots. Studies that have been conducted on group interaction show differences in the perception of a robot in a group, compared to individually. EMP has significant potential for group interaction, through improvement of current issues facing group HRI. These include low willingness to interact as well as higher levels of fear [239]. These issues are often exaggerated for non-anthropomorphic robots in groups, with results indicating that such robots are more threatening and less likely to encourage human engagement [68]. We believe the metrics that often suffer in group interactions can be linked to metrics that have shown improvement through the use of EMP, such as likeability and trust, especially through a reduction of entitativity.

A key issue with groups of robots is the amount of entitativity perceived by human collaborators. Entitativity refers to the level in which a group is seen as a single entity, such as multiple arms being viewed as a single robot, compared to individual agents. Entitativity measures how a group is perceived as a coherent unit rather than separate individuals [240]. Understanding entitativity in human interaction is considered crucial for developing fundamental understandings of human group dynamics [241]. It has been demonstrated that the perception of higher levels of entitativity will create a negative image of the group with less chance of external interaction [242]. For robots, entitativity has only recently entered consideration, with some findings linking higher entitativity to a reduced perception of friendliness and comfort while increasing ratings for “unnervingness” and “creepiness” [243].

In psychology, entitativity is used as an important measurement for group dynamics and effectiveness. Castano theorizes that four main factors impact a group entitativity: common fate, similarity, salience, and boundedness [244]. In human groups, people can relate more to high entitativity groups than low entitativity groups [245, 242]. Hamilton suggests that outsiders are more likely to engage in integrative processing of groups with high entitativity [246]. Increased entitativity will also increase the perceived unification of the group. In human groups, high entitativity requires increased coordination and focus on unification to accomplish a task [246].

While human entitativity has been widely researched, there have been limited studies on the perception of entitativity in robotic groups. Fraune found that increasing the quantity of robots would create more negative emotions towards the robots. A higher quantity would increase anxiety and fear levels of humans [247, 68]. Abrams showed synchronicity in robot movements can vary entitativity and appear scary to an observer. However, robots that appear unique would leave a warmer impression, and increase the desire to work with humans [248]. Saunderson found that a large amount of robots in groups can negatively impact a human’s impression and trust [31]. We believe that the work described in this chapter on integrating EMP into robotic groups can address and mitigate this negative effect of robotic entitativity. Understanding entitativity and HRI metrics in groups leads to our final research question:

RQ 4: Can EMP be scaled to group robotics, to reduce entitativity while increasing trust and likeability ratings?

We believe these benefits can be extended from individual robots to groups of robots, improving key metrics for industrial arms. We also contend that as EMP can be easily modified with timbre shifts, it can support reducing entitativity. Such changes to the sound of the EMP can imply variation between robots in a group setting, allowing an easy format to reduce entitativity.

We conducted a between-groups experiment, comparing three industrial arms performing a collaborative task with a human participant. Participants were shown either the arms without EMP, each arm performing with the same EMP, or the arms performing with variations of the same EMP. We found significant improvements for trust and likeability for the EMP robots, with no variation for participants willingness to interact, confidence with the system or perceived intelligence. Arms performance with different versions of EMP had the lowest rating for entitativity, while using the same EMP achieved higher ratings for entitativity. We also examined the relationship between entitativity ratings and each HRI metric such as trust and likeability, and found that higher levels of entitativity lead to increased ratings across all metrics, contradicting past findings [68].

6.1 Experiment

6.1.1 Method

We investigate three research questions to study the intersection of robots in groups, entitativity and EMP:

RQ 1 Can EMP improve Likeability, Perceived Intelligence, Trust, Confidence and Willingness to Interact, for a group of robots?

RQ 2 Can variations in EMP lower the level of entitativity for a group of robots?

RQ 3 How does the level of entitativity correlate with Likeability, Perceived Intelligence, Trust, Confidence and Willingness to interact?

Research question 1 focuses on understanding the relationship between common HRI metrics and groups of robots. For this question, we are only interested in comparing the same prosodic voice for each robot against gestures, with the goal of replicating improvements shown in past studies with individual robots. The metrics were chosen due to past use in both group studies [68, 239] and the use in our previous studies with individual robots and EMP. Our hypothesis is that each metric will be improved by EMP with a significant result, replicating the results that have occurred for individual robots.

Research question 2 aims to compare the level of entitativity between three groups, one with gestures alone, one with a single voice and one with variations on EMP. Our hypothesis is that the single voice and gesture will perform similarly, while the multiple voices will achieve a lower level of entitativity, implying the appearance of multiple agents in the group.

Research question 3 is an exploratory question, designed to identify the relationship between entitativity and each metrics. We believe that higher levels of entitativity will correlate with reduced metrics as supported by research in human psychology and past research in HRI.

6.1.2 Measures

For each metric we used either an established measure or a combination of existing measures. To measure likeability and perceived intelligence we used a subset of the Godspeed survey [63], as used in chapters 4 and 5. Participants were asked to rate their impression of likeability and perceived intelligence for five questions on a scale of 1-5. We measured willingness to interact and confidence to interact each with three questions on a Likert scale, combined from past surveys [67, 247, 68]. To measure trust we used Schaefer's 14-point scale with participants rating each question from 0-100% to give a total trust percentage. To the common survey answers we added a "Not Applicable" option, as suggested by Chita-Tegmark et al. [249] to allow participants to avoid responding to aspects of trust they feel are not applicable to the industrial arms. We collected participant's age, identified gender and country of origin.

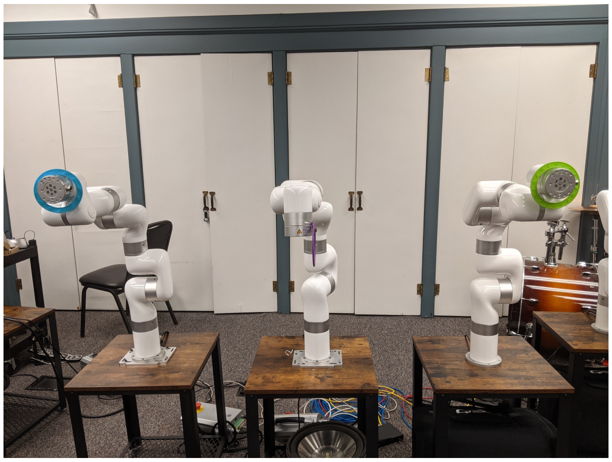


Figure 6.1: Three xArms used for the stimuli. Each xArm was tasked to transport a ring to a box behind them

There is no standard accepted measure of entitativity, with HRI studies commonly combining multiple metrics from social studies, psychology and other HRI papers [250, 247]. Common questions range from defining entitativity for the participants and then asking directly for a rating [251], to attempts to combine other metrics such as friendliness, creepiness, comfort and unnerving into a rating [243]. We chose to measure entitativity using the survey proposed and validated by Blanchard et al. [241] which was shown to be effective for online and in person analysis. This measure consisted of three questions on a 7-point Likert scale.

Stimuli

We used a two minute video as our stimuli, overdubbed with different audio for each group. In the video we showed three robotic arms (shown in Figure 6.1) interacting with a human user. The human user placed a ring on each arm, that the robot then placed in a box behind itself. Each robot used the same movement and gestures to place the rings in the box. We chose to have each robot act directly with the human, as opposed to additional interactions from robot to robot. While this limited the group dynamic, we believed this allowed a better experimental setup for multiple reasons. This allowed identical performance and interaction for each robot, with the same repeated action. It also allowed for more opportunities for the human to interact with the robot and hear the EMP response. Finally, we believed by containing to a simple interaction we could highly control users perception of the interaction between response types. The robot used in the study was an xArm, a 7 degree of freedom industrial arm made by uFactory.

We created three versions of the video with different audio, starting with a gesture only version which did not have any added audio. From the previously generated audio we chose emotions tagged as admiration, contentment, and compassion, each low arousal high valence emotions.

The second version of the video used a matching voice (referred henceforth as single voice) for each robot. For each interaction the single voice used a different prosodic phrase, but had matching timbre, essentially sounding like the same voice singing a different phrase each time. For the third version of the video we used three different versions of the voice from the dataset. We also added variations to each voice through pitch shifting, a formant filter and modulation. This had the effect of sounding like three different voices, one for each robot. All three versions maintained the room sound and sounds of the robots movements. All stimuli can be viewed online. [1]

6.1.3 Participants

We recruited 60 participants on Prolific and 108 participants on Amazon Mechanical Turk (MTurk) to complete the study. Each participant was paid $2.00. We selected only MTurk Masters to participate and had no restrictions on prolific. We used multiple attention checks to verify each participant, and disqualified any data that failed any check. Our first attention check consisted of a spoken phrase at the end of the video requesting participants to type a random word on the next screen. We also had a question in the trust survey requiring participants to choose 10%. In addition to direct questions, we tracked the time spent on each question and the video, with any participant who did not watch the entire video

[1] www.richardsavery.com/prosodyentatitivitystudy

removed. Finally, we removed two participants who completed the survey a second time, we assume after realizing they missed the audio from the attention check and restarting. From Prolific 6 participants failed an attention check, while 9 on MTurk failed an attention check, leaving us with a total of 153 participants.

In total we had 49 participants in the gesture only group, 48 in single EMP and 56 in the multiple audio. Participants place of origin was spread across 22 countries, with the majority from United States of America (n=71), India (n=22), Poland (n=14), Portugal (n=11), Mexico (n=8) with the remaining countries each have 5 or less participants. We found no significant variation in responses from each country, with the countries with less than 5 each fitting within the range of majority of responses. We had 62 participants identify as female and 90 as male, also with no significant variation between groups. The mean age of participants was 37 with a standard deviation of 12 and ranging from 18 to 75.

6.1.4 Protocol

The survey was conducted online using Qualtrics. Participants first completed a consent form and entered their MTurk or Prolific ID to indicate consent. They were then given instructions to watch the stimuli video with headphones connected. Participants were randomly assigned to one of the three groups of the study. Following the video participants first entered the text for the attention check and then completed the previously described measures. The measures were randomly ordered for each participant, with the sub-questions (such as each component of the trust survey) also randomly ordered for each participants. After completing each measure participants entered their demographic details and had a open text field with a prompt asking for any feedback on the robot system or experiment in general.

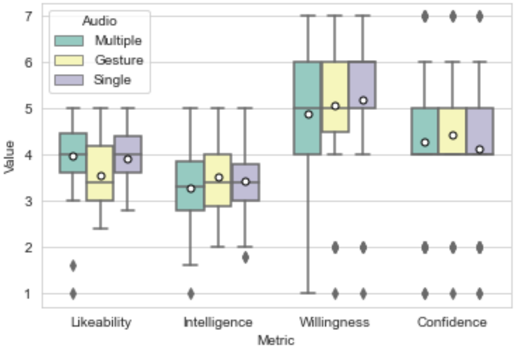


Figure 6.2: Box Plot of Likeability, Intelligence, Willingness and Confidence

6.2 Results

6.2.1 RQ 1: HRI Metrics

Likeability and Perceived Intelligence

The Cronbach's Alpha results for Likeability and Perceived Intelligence were 0.869 and 0.866 respectively, indicating high internal reliability for both measures. Perceived Intelligence had the results for single voice (mean = 3.324, std = 0.716, effect size = 0.050), multiple voices (mean = 3.271, std = 0.880, effect size = 0.230) and the gestures alone (mean = 3.527, std = 0.764, effect size = 0.240), with effect size calculated using Cohen's D. We ran a one-way ANOVA with the result $p > 0.05$, indicating the result was not significant. Perceived intelligence did not have a significant different between groups with each category having similar means and standard deviations, which did not support our hypothesis.

Likeability had the results for single voice (mean = 3.931, std = 0.600, effect size = 0.246), multiple voices (mean = 3.975, std = 0.800, effect size = 0.285) and the gestures alone (mean = 3.553, std = 0.736, effect size = 0.545), with effect size calculated using Cohen's D. We ran a one-way ANOVA with the result p = 0.007, indicating the result was significant. Likeability was improved significantly for both versions of EMP over the gestures alone, supporting our hypothesis. Figure 6.2 shows a box plot of the results for likeability and perceived intelligence. Perceived intelligence did not have a significant different between groups with each category having similar means and standard deviations, which did not support our hypothesis.

Confidence and Willingness

Confidence had the results for single voice (mean = 4.142, std = 1.607, effect size = 0.128), multiple voices (mean = 4.285, std = 1.637, effect size = 0.003) and the gestures alone (mean = 4.42, std = 1.363, effect size = 0.151), with effect size calculated using Cohen's D. We ran a one-way ANOVA with the result $p > 0.05$, indicating the result was not significant. Willingness had the results for single voice (mean = 5.183, std = 1.409, effect size = 0.158), multiple voices (mean = 4.875, std = 1.663, effect size = 0.150) and the gestures alone (mean = 5.064, std = 1.699, effect size = 0.026), with effect size calculated using Cohen's D. We ran a one-way ANOVA with the result $p > 0.05$, indicating the result was not significant. Neither confidence or willingness showed a significant result, indicating that EMP did not improve either of these metrics. Figure 6.2 shows a box plot of these results.

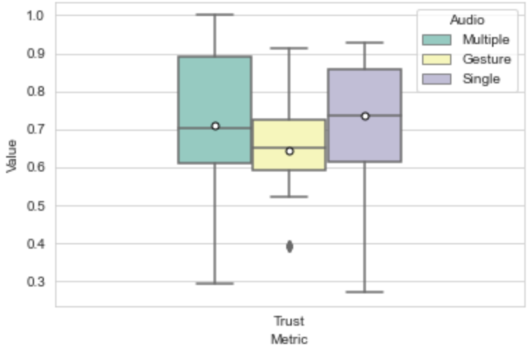


Figure 6.3: Box Plot of Trust Ratings

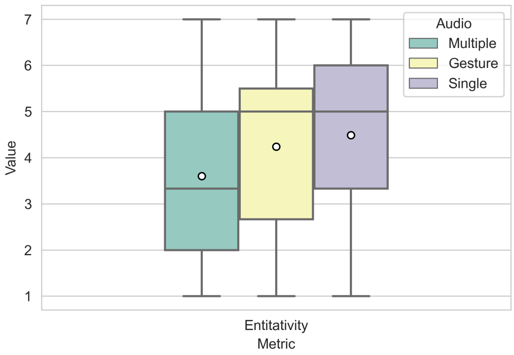


Figure 6.4: Box Plot of Entitativity Ratings

Trust

To analyze our trust results we first calculated Cronbach's alpha which gave the result of 0.859, indicating high internal reliability. For single voice the results were (mean = 0.734, std = 0.146, effect size = 0.376), multiple voices (mean = 0.710, std = 0.166, effect size = 0.125) and the gestures alone (mean = 0.642, std = 0.710, effect size = 0.592), with effect size calculated using Cohen's D. We ran a one-way ANOVA with the result p = 0.009, indicating the result was significant. This supported our hypothesis that EMP would increase trust over gesture. Figure 6.3 shows the results as a box plot.

6.2.2 RQ 2: Entitativity and EMP

For the three entitativity questions we first calculated Cronbach's Alpha, which gave a result of 0.88, indicating high internal reliability across the questions. For gestures alone the results were (mean = 4.241, std = 1.699), the single voice (mean = 4.490, std = 1.667) and multiple robots (mean = 3.601, std = 1.706). A one-way ANOVA gave a p-value of 0.022 indicating the results was significant. Additionally the multiple voices had an effect size calculated with Cohen's D of 0.45, indicating a medium effect size. This supported our hypothesis that subtle variations in voice would increase the entitativity of the group. Figure 6.4 shows a box plot of the results.

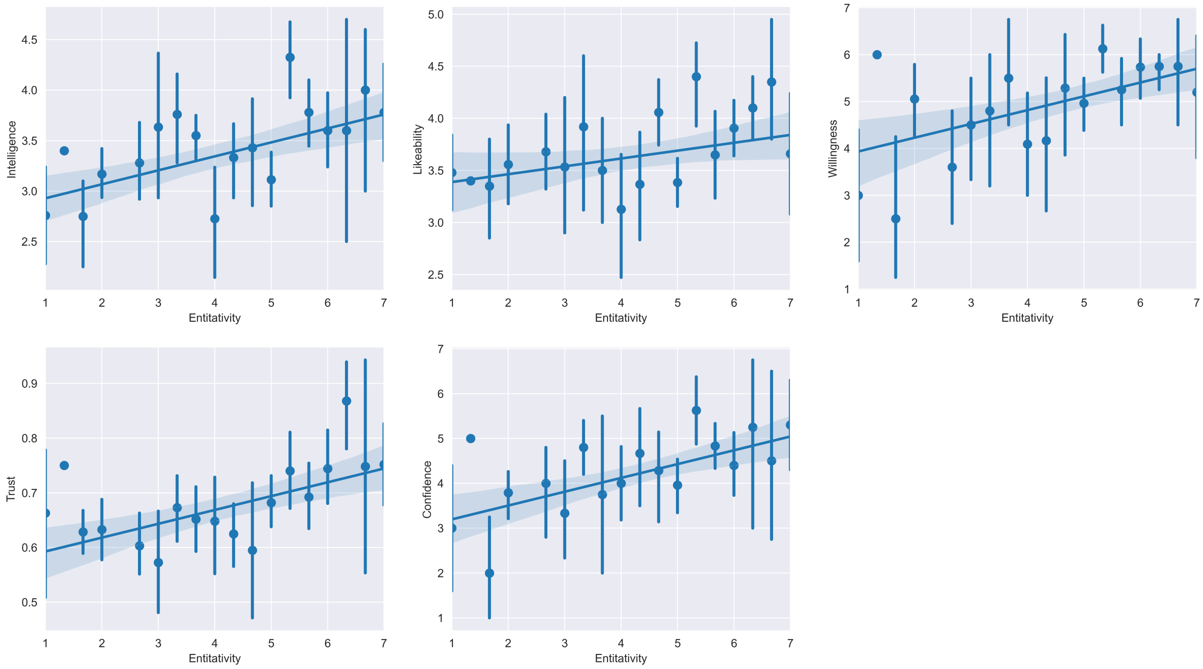


Figure 6.5: Linear Regression comparing entitativity with HRI Metrics

6.2.3 RQ 3: Entitativity and HRI Metrics

For research question 3 we fit a linear regression model for each metric with entitativity. Table 6.1 shows the slope, intercept, r, p and error for each metric. Each metric tested had a positive slope, with higher levels of entitativity correlating with higher ratings. This did not support our hypothesis, as we had expected the opposite to occur across every metric.

Table 6.1: Linear Regression Statistics

	Slope	Intercept	r	p	Error
Willingness	0.294	3.64	0.286	p <.001	0.081
Intelligence	0.138	2.79	0.29	p <.001	0.037
Trust	0.025	0.567	0.31	p <.001	0.006
Likeability	0.075	3.313	0.162	0.046	0.037
Confidence	0.308	2.888	0.327	p <.001	0.073

6.3 Discussion

6.3.1 RQ 1: HRI Metrics

We found that embedding EMP to accompany co-bot arms gestures improved human's trust and likeability for these robots with significant results. Since in previous work, EMP improved trust and likeability in individual robots, it was expected that the improvements would carry across to groups. These metrics supports one of the core principles behind the use of EMP in robots, namely that by increasing a robot's presence as an engaging emotional agent, human's will trust it and like to interact with it more. Since these results occurred for both versions of EMP, we propose that these metrics are relatively robust to variations in timbre and EMP.

In previous chapters, embedding EMP in robotic actions has been shown to increase perceived intelligence for individual robots. However in those studies the interactions were more social in nature, either with a social robot or a collaborative pattern recognition task with the arm. In our current experiment, where the robot was expected to perform a task (moving rings and placing them in a box) we propose that the successful performance by the robot was more influential on users' perception of its intelligence than external factors such as EMP.

Our initial hypothesis that willingness and confidence would be improved with EMP was not supported. In past work the effect of EMP has not yet been used on individual robots for these two factors. It is not clear from this study whether EMP can influence these metrics which requires future study.

6.3.2 RQ 2: Entitativity and EMP

Our results for research question 2 indicated that multiple voices did lower entitativity, increasing the perception of the group of robots as individual agents. This increase was achieved with only subtle variations, that could be easily achieved in real-time and scaled to many robots. We did not predict that having a single voice would increase entitativity however, as believed the gestures alone would appear as a group and single EMP would maintain this level. This reflected our original belief that entitativity would be relatively insusceptible to being increased amongst robots that already look and move in an identical manner. This finding has future implications for the possibility of audio design to not only reduce entitativity as per our original goal, but also the possibility of raising the level of entitativity.

6.3.3 RQ 3: Entitativity and HRI Metrics

A key finding in this study was the relation between entitativity and common HRI metrics. Our findings differ from those of related work on robots and groups [68]. This correlation between higher entitativity and each metric occurred across all groups independently, with gestures, single voice and multiple voice all showing the same relationship. We believe extensive future research should be undertaken to establish more completely the relationship between entitativity and groups of robots. We suggest that a possible explanation may be that with each robot performing the same task, participants may generally prefer interacting with the robot when perceived as a single agent, rather than having to engage with multiple agents. Multiple robots performing a similar task could give a perception that the robots are uniting towards a common goal. This would give participants a more positive impression that the robots are likeable and cooperative. This explanation would match Hamilton's studies on human groups that outsiders are more likely to engage with groups that have a higher entitativity [246].

6.3.4 Sound for Functionality

The majority of subjects text responses ranged from one to four sentences, and generally did not show much variation between groups. One standout comment was that 8 participants from both EMP groups commented that they were not sure what the purpose of the sounds was. One participant noted: “I thought the singing was interesting but I don’t see how that relates to the task success of the robot”. Despite recognizing that the audio was not functional in the clip, this participant’s ratings were well above the mean for each category, and we saw no reduction in ratings for any participant who noted there was no functional purpose. Nevertheless, we believe there is significant possibilities in considering the different applications of functional compared to non-functional or auxiliary sound and understanding how that impacts individual as well as groups of robots.

6.3.5 Limitations

This study was performed online using pre-recorded videos instead of live interaction or video watching in person. We believe that for this experiment this was an acceptable experimental design as ultimately our analysis focused on external viewing and analyzing a group of robots. Multiple past papers have shown no significant variation in results when a participant is watching a robot on video compared to in person [200, 201]. We also believe the use of MTurk and Prolific has some advantages over in person studies, allowing us a far larger and more diverse participant pool than possible in person. It has also been shown that compared to university pools, MTurk participants are more careful [252]. When combined with our multiple point attention check we are confident that our results would be replicated in person.

We chose to use an industrial arm as they are commonly used in group manufacturing settings. In future studies we are interested in researching how the impact of EMP on robotic groups varies between platforms such as social or humanoid robots. Likewise, we only compared EMP to no audio, and in the future expect to compare different audio conditions.

6.4 Conclusion

In this chapter we were able to show with significant results, that EMP improves likeability and trust when used in a group of three industrial robots. We also showed that variations in EMP can lead to lower levels of entitativity, however a single voice can raise the level of entitativity. Our results analyzing the correlation between entitativity and other HRI metrics suggest a wide-range of future research to understand the wider impact entitativity has on collaborative robots.

CHAPTER 7
CONCLUSION

In this dissertation, we have presented a new method for human-robot communication built on emotional musical prosody (EMP). In this final chapter, a summary of the results and contributions is presented, followed by potential future directions and final remarks. Figure 7.1 displays an overview of the dissertation structure. After first developing the dataset and EMP generator, we next studied the applications across multiple platforms. We showed significant results for the social robot, robot musician and industrial arm, consistently improving trust and likeability. We then focused on the robotic arm, due to its widespread adoption and our belief that EMP can most improve issues arising from the habitability gap in this platform. This was followed by the development of personality interactions, to explore a framework for robots to autonomously use emotion in their interactions. Finally, we explored how EMP could function in a group of robotic arms.

7.1 Research Questions and Contributions

We found significant results for each research question. Table 7 shows an overview of the robotic platforms used. The table also includes the metrics studied, the type of interaction analyzed, what audio and gestures were compared, the significant results, and broad findings.

7.1.1 EMP Generation and Dataset Collection

RQ 1: How can a data driven, EMP system generate musical phrases that can be labelled by listeners?

After creating a new dataset, the EMP generative system was able to generate new phrases in real-time for robot interaction. The generated EMP phrases were labelled by

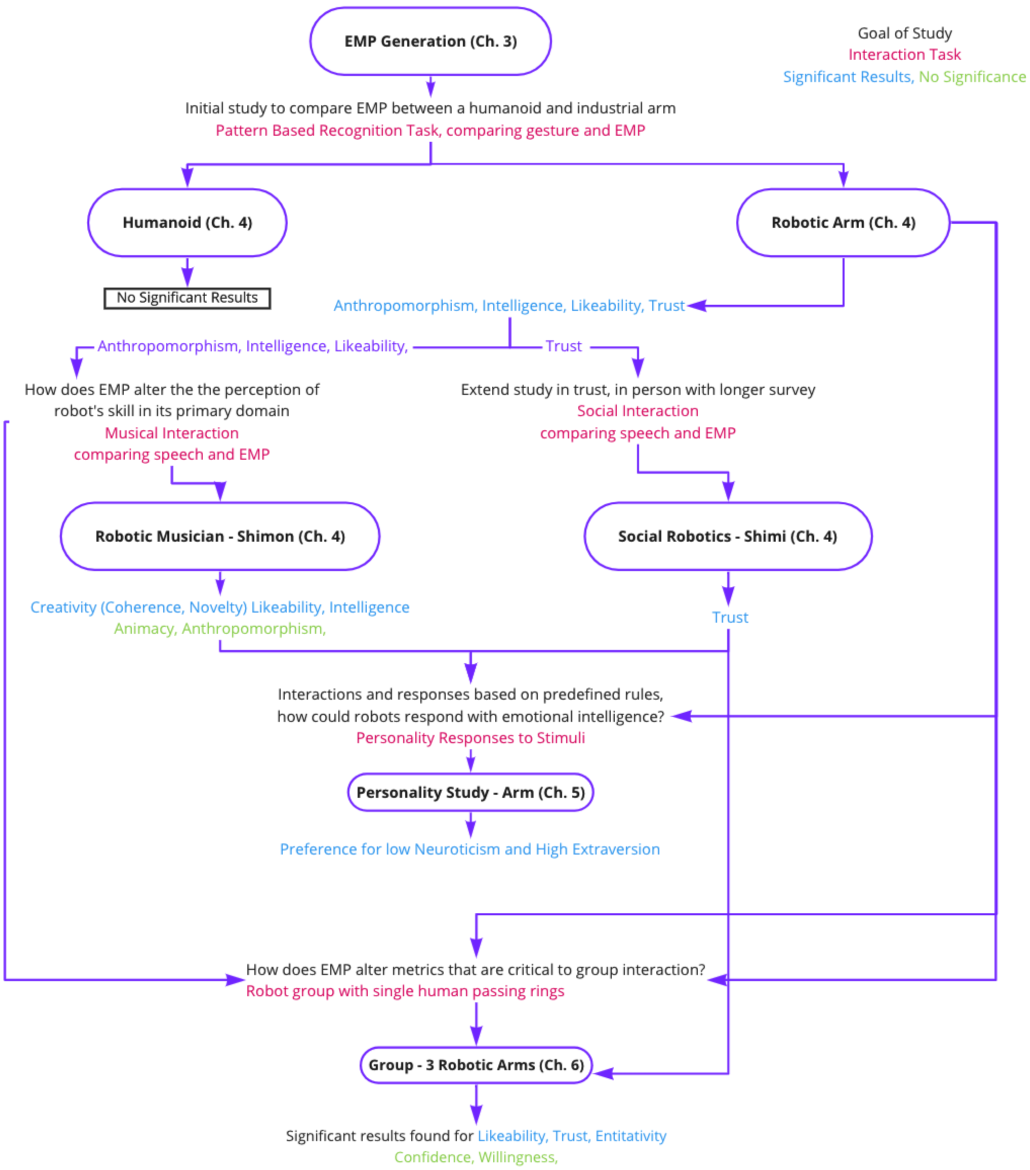

Goal of Study
Interaction Task
Significant Results, No Significance
EMP Generation (Ch. 3)
Initial study to compare EMP between a humanoid and industrial arm
Pattern Based Recognition Task, comparing gesture and EMP
Humanoid (Ch. 4)
Robotic Arm (Ch. 4)
No Significant Results
Anthropomorphism, Intelligence, Likeability, Trust
Anthropomorphism, Intelligence, Likeability,
Trust
How does EMP alter the the perception of robot's skill in its primary domain
Musical Interaction
comparing speech and EMP
Extend study in trust, in person with longer survey
Social Interaction
comparing speech and EMP
Robotic Musician - Shimon (Ch. 4)
Social Robotics - Shimi (Ch. 4)
Creativity (Coherence, Novelty) Likeability, Intelligence
Animacy, Anthropomorphism,
Trust
Interactions and responses based on predefined rules, how could robots respond with emotional intelligence?
Personality Responses to Stimuli
Personality Study - Arm (Ch. 5)
Preference for low Neuroticism and High Extraversion
How does EMP alter metrics that are critical to group interaction?
Robot group with single human passing rings
Group - 3 Robotic Arms (Ch. 6)
Significant results found for Likeability, Trust, Entitativity
Confidence, Willingness,

Figure 7.1: Dissertation Structure

humans with an accuracy similar to that of the original EMP dataset. This research question enabled two primary contributions, firstly the creation of the dataset itself. This dataset not only enabled the creation of the generative system, but also showed the validity of computer and human understanding of emotion in musical prosody. The second contribution was the development of a new model for EMP generation, demonstrating a use case for a CVAE in generating classifiable musical emotions.

7.1.2 EMP, Trust and HRI Metrics

RQ 2: How can EMP alter the level of likeability, perceived intelligence and trust in social, industrial, humanoid and musical robots?

We were able to demonstrate improvement in multiple metrics, across each platform except humanoid robots. Table 7 shows a full summary of results. Overall this research question confirmed our novel approach of building trust using EMP to communicate with human collaborators. We were able to confirm our hypothesis that by avoiding the uncanny valley and habitability gap we could improve metrics such as trust and likeability. Other key takeaways from this research question include that EMP did not reduce ratings for intelligence when compared to speech, not only suggesting the validity of EMP as an approach but also indicating that intelligence is not tied to speech in robotics.

7.1.3 Personality Preferences

RQ 3: Does a person's level of neuroticism and extraversion affect their ratings of different emotional responses portrayed through EMP?

In this research question we developed new knowledge about humans' preference for robotic emotional response based on personality, based on neuroticism and extraversion traits. The use of personality in this dissertation presents some early steps towards leveraging personality for interaction. We believe this presents previously unexplored options for research in robot customization, based on human personality type. This can lead to implication not only in embedding audio features in robots as described in this paper, but also consideration of all areas of robotic design. By embedding personality traits in robots through design variations we believe robotics can be better developed for specific interactions and human experiences. Finally, we believe this research outlines the need for further consideration and research of personality traits and their links to human robot interaction.

7.1.4 Group Interaction

RQ 4: How can EMP be leveraged in a group of three identical robots and one human participant, to reduce entitativity while

We were able to show that by using EMP in a group interaction we were able to improve likeability and trust. Through subtle variations in timbre we were also able to reduce the level of entitativity. Finally, our results indicate a complex relationship between entitativity and common HRI metrics with higher levels of entitativity leading to improved performance, contradicting past literature.

7.1.5 Music Technology

In addition to the core findings from each research question, this dissertation also presents significant contributions to music technology. We believe firstly that with extra study the newly developed dataset of EMP can provide new knowledge about the use of and portrayal of emotion in music, as well as providing many future opportunities for both musicological and machine learning research. This work also shows the potential for sound to drastically alter the perception of technology, potentially informing the design of digital musical instruments. Most importantly this work demonstrates how creative and music technology driven approaches can further permeate and inform the future of technology.

7.2 Future Work

7.2.1 Long-term Studies

Like many HRI applications our experiments only occurred over a small time frame and did not consider long-term implications of the system [253]. The use of EMP has not yet been studied in long-term applications, but may have different use cases and would require additional changes in the implementation. One participant from the group and entitativity study commented on the time scale, describing: “I really like it in the short term but I feel like I'd get tired of it if I had to listen all day long". In the future we are interested in applying EMP to longer form interactions in person and considering how EMP can be adjusted for use not across many sessions.

7.2.2 Audio in HRI

We believe this work indicates the importance of audio design in robotics, and the impact that robotic audio can have on human perception. Through changing audio alone and not relying on default audio methods such as speech, we were able to drastically change the perception of a robotic system. While we have shown EMP as particularly effective at improving a range of metrics, this is just one of the many possible approaches that could be developed with more careful future audio design for human-robot interaction

7.3 Final Remarks

At the core of this research is the development of the vastly under-explored potential for musical approaches to drive interaction with artificial intelligence and robotics. We were able to demonstrate that through an implementation of EMP we could show significant improvement in a range of metrics for robot interaction. As we have shown throughout the dissertation, EMP as an audio technique is uniquely positioned to leverage the advantages of musical and human communication for improved human-robot interaction.

Table 7.1: Summary of Dissertation Findings

Ch.	Robot	DOF	Metrics	Interaction Type	Comparing	Significant Results	Broad findings
4	Arm (Simulation)	4	Anthropomorphism, Intelligence, Likeability, Trust	Pattern recognition	EMP, gestures, non-prosody audio	Trust, pairwise between Godspeed	Relation between trust ratings and user choices
4	Humanoid (Simulation)	18	Anthropomorphism, Intelligence, Likeability, Trust	Pattern recognition	EMP, gestures, non-prosody audio	None	Relation between trust ratings and user choices
4	Social Robot (Shimi)	5	Trust	Social Interaction	EMP, Speech	Trust	Speech can reduce trust
4	Robotic Musician (Shimon)	5	Creativity (Coherence, Novelty, Expressivity), Animacy, Anthropomorphism, Likeability, Intelligence	Musical improvisation	EMP, Speech	Creativity (Coherence, Novelty) Likeability, Intelligence	Audio outside core task can alter perception of key task
5	Arm (Panda)	7	Animacy, Anthropomorphism, Likeability, Intelligence	Personality response to emotion tagged visuals	EMP prefences by Neuroticism and Extraversion Traits	Preference for low Neuroticism and High Extraversion	Human personality type preferences for robotic platforms
6	Arm (xArm)	7	Likeability, Intelligence, Confidence, Willingness, Trust, Entitativity	Group interaction, 3 robots	EMP, EMP with timbre changes, gestures	Likeability, Trust, Entitativity	Entitativity is not correlated to other metrics as previously proposed

REFERENCES

[1] R. K. Moore, “Is spoken language all-or-nothing? implications for future speech-based human-machine interaction,” in *Dialogues with Social Robots*, Springer, 2017, pp. 281–291.

[2] ——, “Appropriate voices for artefacts: Some key insights,” in *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, 2017.

[3] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, “A meta-analysis of factors affecting trust in human-robot interaction,” *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[4] M. Mateas, “Artificial intelligence today,” in, M. J. Wooldridge and M. Veloso, Eds., Berlin, Heidelberg: Springer-Verlag, 1999, ch. An Oz-centric Review of Interactive Drama and Believable Agents, pp. 297–328, ISBN: 3-540-66428-9.

[5] J. Bates, “The role of emotion in believable agents,” *Commun. ACM*, vol. 37, no. 7, pp. 122–125, Jul. 1994.

[6] R. A. Khan and J. Chitode, “Concatenative speech synthesis: A review,” *International Journal of Computer Applications*, vol. 136, no. 3, p. 6, 2016.

[7] P. Ekman, “Facial expression and emotion.,” *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[8] T. Hashimoto, S. Hitramatsu, T. Tsuji, and H. Kobayashi, “Development of the face robot saya for rich facial expressions,” in *2006 SICE-ICASE International Joint Conference*, IEEE, 2006, pp. 5423–5428.

[9] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[10] A. Wennerstrom, *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press, 2001.

[11] C. Palmer and S. Hutchins, “What is musical prosody?” *Psychology of Learning and Motivation*, vol. 46, Dec. 2006.

[12] J. Crumpton and C. L. Bethel, "A survey of using vocal prosody to convey emotion in robot speech," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 271–285, 2016.

[13] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Autonomous robots*, vol. 12, no. 1, pp. 83–104, 2002.

[14] J. Sloboda, "Music: Where cognition and emotion meet," in *Conference Proceedings: Opening the Umbrella; an Encompassing View of Music Education; Australian Society for Music Education, XII National Conference, University of Sydney, NSW, Australia, 09-13 July 1999*, Australian Society for Music Education, 1999, p. 175.

[15] W. F. Thompson, E. G. Schellenberg, and G. Husain, "Decoding speech prosody: Do music lessons help?" *Emotion*, vol. 4, no. 1, p. 46, 2004.

[16] M. Hausen, R. Torppa, V. R. Salmela, M. Vainio, and T. Särkämö, "Music and speech prosody: A common rhythm," *Frontiers in psychology*, vol. 4, p. 566, 2013.

[17] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, p. 715, 2005.

[18] B. Hendriks, B. Meerbeek, S. Boess, S. Pauws, and M. Sonneveld, "Robot vacuum cleaner personality and behavior," *International Journal of Social Robotics*, vol. 3, no. 2, pp. 187–195, 2011.

[19] W. Revelle and K. R. Scherer, "Personality and emotion," *Oxford companion to emotion and the affective sciences*, vol. 1, pp. 304–306, 2009.

[20] U. Barańczuk, "The five factor model of personality and emotion regulation: A meta-analysis," *Personality and Individual Differences*, vol. 139, pp. 217–227, 2019.

[21] J. A. Penley and J. Tomaka, "Associations among the big five, emotional responses, and coping with acute stress," *Personality and individual differences*, vol. 32, no. 7, pp. 1215–1228, 2002.

[22] B. Parkinson, "Emotions are social," *British journal of psychology*, vol. 87, no. 4, pp. 663–683, 1996.

[23] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, pp. 22–35, 2015.

[24] K. Dautenhahn, M. Walters, S. Woods, K. L. Koay, C. L. Nehaniv, A. Sisbot, R. Alami, and T. Siméon, "How may i serve you? a robot companion approaching a

seated person in a helping context," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, ser. HRI '06, Salt Lake City, Utah, USA: Association for Computing Machinery, 2006, 172–179, ISBN: 1595932941.

[25] J. N. Pires and A. S. Azar, "Advances in robotics for additive/hybrid manufacturing: Robot control, speech interface and path planning," *Industrial Robot: An International Journal*, 2018.

[26] J. Cambre, J. Colnago, J. Maddock, J. Tsai, and J. Kaye, "Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[27] J. Cambre and C. Kulkarni, "One voice fits all? social implications and research challenges of designing voices for smart devices," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Nov. 2019.

[28] B. R. Cowan, N. Pantidi, D. Coyle, K. Morrissey, P. Clarke, S. Al-Shehri, D. Earley, and N. Bandeira, ""what can i help you with?": Infrequent users' experiences of intelligent personal assistants," in *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, ser. MobileHCI '17, Vienna, Austria: Association for Computing Machinery, 2017, ISBN: 9781450350754.

[29] M. P. Aylett, S. J. Sutton, and Y. Vazquez-Alvarez, "The right kind of unnatural: Designing a robot voice," in *Proceedings of the 1st International Conference on Conversational User Interfaces*, ser. CUI '19, Dublin, Ireland: Association for Computing Machinery, 2019, ISBN: 9781450371872.

[30] R. Jones, *Communication in the real world: An introduction to communication studies*. The Saylor Foundation, 2013.

[31] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *International Journal of Social Robotics*, vol. 11, no. 4, pp. 575–608, 2019.

[32] B. Gleeson, K. MacLean, A. Haddadi, E. Croft, and J. Alcazar, "Gestures for industry intuitive human-robot communication from human observation," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2013, pp. 349–356.

[33] M. L. Walters, K. Dautenhahn, R. Te Boekhorst, K. L. Koay, C. Kaouri, S. Woods, C. Nehaniv, D. Lee, and I. Werry, "The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment," in *ROMAN 2005.*

IEEE International Workshop on Robot and Human Interactive Communication, 2005., IEEE, 2005, pp. 347–352.

[34] H. Fukuda, M. Shiomi, K. Nakagawa, and K. Ueda, “'midas touch'in human-robot interaction: Evidence from event-related potentials during the ultimatum game,” in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2012, pp. 131–132.

[35] A. Moon, C. A. Parker, E. A. Croft, and H. M. Van der Loos, “Did you see it hesitate?-empirically grounded design of hesitation trajectories for collaborative robots,” in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2011, pp. 1994–1999.

[36] J. Goetz, S. Kiesler, and A. Powers, “Matching robot appearance and behavior to tasks to improve human-robot cooperation,” in *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. RO-MAN 2003.*, Ieee, 2003, pp. 55–60.

[37] M. S. Erden, “Emotional postures for the humanoid-robot nao,” *International Journal of Social Robotics*, vol. 5, no. 4, pp. 441–456, 2013.

[38] E. Rosen, D. Whitney, E. Phillips, G. Chien, J. Tompkin, G. Konidaris, and S. Tellex, “Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays,” *The International Journal of Robotics Research*, vol. 38, no. 12-13, pp. 1513–1526, 2019.

[39] E. Cha, Y. Kim, T. Fong, M. J. Mataric, *et al.*, “A survey of nonverbal signaling methods for non-humanoid robots,” *Foundations and Trends® in Robotics*, vol. 6, no. 4, pp. 211–323, 2018.

[40] C. Bodden, D. Rakita, B. Mutlu, and M. Gleicher, “Evaluating intent-expressive robot arm motion,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2016, pp. 658–663.

[41] E. Ruffaldi, F. Brizzi, F. Tecchia, and S. Bacinelli, “Third point of view augmented reality for robot intentions visualization,” in *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, Springer, 2016, pp. 471–478.

[42] S. Yilmazyildiz, R. Read, T. Belpeame, and W. Verhelst, “Review of semantic-free utterances in social human–robot interaction,” *International Journal of Human-Computer Interaction*, vol. 32, no. 1, pp. 63–85, 2016.

[43] K. Fischer, K. Lohan, J. Saunders, C. Nehaniv, B. Wrede, and K. Rohlfing, "The impact of the contingency of robot feedback on hri," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, IEEE, 2013, pp. 210–217.

[44] M. R. Frederiksen and K. Stoey, "Augmenting the audio-based expression modality of a non-affective robot," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019, pp. 144–149.

[45] E. Cha, N. T. Fitter, Y. Kim, T. Fong, and M. J. Matarić, "Effects of robot sound on auditory localization in human-robot collaboration," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 434–442.

[46] H. Tennent, D. Moore, M. Jung, and W. Ju, "Good vibrations: How consequential sounds affect perception of robotic arms," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2017, pp. 928–935.

[47] D. Moore, N. Martelaro, W. Ju, and H. Tennent, "Making noise intentional: A study of servo sound perception," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, IEEE, 2017, pp. 12–21.

[48] R. Zhang, J. Barnes, J. Ryan, M. Jeon, C. H. Park, and A. Howard, "Musical robots for children with asd using a client-server architecture," in *International Conference on Auditory Display*, 2016.

[49] J. Bellona, L. Bai, L. Dahl, and A. LaViers, "Empirically informed sound synthesis application for enhancing the perception of expressive robotic movement," Georgia Institute of Technology, 2017.

[50] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[51] M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th australasian conference on information systems*, Citeseer, vol. 53, 2000, pp. 6–8.

[52] N. Moray and T Inagaki, "Laboratory studies of trust between humans and machines in automated systems," *Transactions of the Institute of Measurement and Control*, vol. 21, no. 4-5, pp. 203–211, 1999.

[53] P. H. Kim, K. T. Dirks, and C. D. Cooper, "The repair of trust: A dynamic bilateral perspective and multilevel conceptualization," *Academy of Management Review*, vol. 34, no. 3, pp. 401–422, 2009.

[54] R. E. Miles and W. D. Creed, “Organizational forms and managerial philosophies-a descriptive and analytical review,” *RESEARCH IN ORGANIZATIONAL BEHAVIOR: AN ANNUAL SERIES OF ANALYTICAL ESSAYS AND CRITICAL REVIEWS, VOL 17, 1995*, vol. 17, pp. 333–372, 1995.

[55] K. E. Schaefer, “Measuring trust in human robot interactions: Development of the “trust perception scale-hri”,” in *Robust Intelligence and Trust in Autonomous Systems*, R. Mittu, D. Sofge, A. Wagner, and W. Lawless, Eds. Boston, MA: Springer US, 2016, pp. 191–218, ISBN: 978-1-4899-7668-0.

[56] P. M. Satchell, *Cockpit monitoring and alerting systems*. Routledge, 2016.

[57] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, “Measurement of trust in human-robot collaboration,” in *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on*, IEEE, 2007, pp. 106–114.

[58] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, “Not so different after all: A cross-discipline view of trust,” *Academy of management review*, vol. 23, no. 3, pp. 393–404, 1998.

[59] T. Gompei and H. Umemuro, “Factors and development of cognitive and affective trust on social robots,” in *International Conference on Social Robotics*, Springer, 2018, pp. 45–54.

[60] S. G. Barsade, “The ripple effect: Emotional contagion and its influence on group behavior,” *Administrative science quarterly*, vol. 47, no. 4, pp. 644–675, 2002.

[61] R. M. Stock, “Emotion transfer from frontline social robots to human customers during service encounters: Testing an artificial emotional contagion modell,” 2016.

[62] S. Ososky, D. Schuster, E. Phillips, and F. G. Jentsch, “Building appropriate trust in human-robot teams,” in *2013 AAAI Spring Symposium Series*, 2013.

[63] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots,” *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[64] A. Kaplan, T. Sanders, and P. Hancock, “Likert or not? how using likert rather than biposlar ratings reveal individual difference scores using the godspeed scales,” *International Journal of Social Robotics*, pp. 1–10, 2021.

[65] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, “The robotic social attributes scale (rosas) development and validation,” in *Proceedings of the 2017*

ACM/IEEE International Conference on human-robot interaction, 2017, pp. 254–262.

[66] A. Weiss and C. Bartneck, “Meta analysis of the usage of the godspeed questionnaire series,” in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2015, pp. 381–388.

[67] N. L. Robinson, T.-N. Hicks, G. Suddrey, and D. J. Kavanagh, “The robot self-efficacy scale: Robot self-efficacy, likability and willingness to interact increases after a robot-delivered tutorial,” in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2020, pp. 272–277.

[68] M. R. Fraune, S. Šabanović, E. R. Smith, Y. Nishiwaki, and M. Okada, “Threatening flocks and mindful snowflakes: How group entitativity affects perceptions of robots,” in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, IEEE, 2017, pp. 205–213.

[69] M. N. Kozak, A. A. Marsh, and D. M. Wegner, “What do i think you’re doing? action identification and mind attribution.,” *Journal of personality and social psychology*, vol. 90, no. 4, p. 543, 2006.

[70] M. Gendron, “Defining emotion: A brief history,” *Emotion Review*, vol. 2, no. 4, pp. 371–372, 2010.

[71] P. Ekman, “Basic emotions,” *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16,

[72] D Watson, “How are emotions distinguished from mood, temperament, and other related affective constructs,” *The nature of emotion*, 1994.

[73] R. Savery and G. Weinberg, “A survey of robotics and emotion: Classifications and models of emotional interaction,” in *Proceedings of the 29th International Conference on Robot and Human Interactive Communication*, 2020.

[74] R. C. Arkin and P. Ulam, “An ethical adaptor: Behavioral modification derived from moral emotions,” in *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA)*, IEEE, 2009, pp. 381–387.

[75] R. C. Arkin, M. Fujita, T. Takagi, and R. Hasegawa, “An ethological and emotional basis for human–robot interaction,” *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 191–201, 2003.

[76] T. Ogata and S. Sugano, "Emotional communication robot: Wamoeba-2r emotion model and evaluation experiments," in *Proceedings of the International Conference on Humanoid Robots*, 2000.

[77] L. Cañamero and R. Aylett, *Animating expressive characters for social interaction*. John Benjamins Publishing Company, 2008, vol. 74.

[78] A. J. Fridlund, P. Ekman, and H. Oster, "Facial expressions of emotion.," 1987.

[79] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation," in *Proc. speech prosody*, 2006.

[80] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[81] M. Inderbitzin, A. Väljamäe, J. M. B. Calvo, P. F. M. J. Verschure, and U. Bernardet, "Expression of emotional states during locomotion based on canonical parameters," in *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*, IEEE, 2011, pp. 809–814.

[82] H. G. Walbott, "Bodily expression of emotion," *European Journal of Social Psychology*, vol. 28, no. 6, pp. 879–896, 1998.

[83] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[84] B. de Gelder, "Towards the neurobiology of emotional body language," *Nature Reviews Neuroscience*, vol. 7, pp. 242–249, Mar. 2006.

[85] M. M. Nele Dael and K. R. Scherer, "The body action and posture coding system (bap): Development and reliability," *Journal of Nonverbal Behavior*, vol. 36, pp. 97–121, 2012.

[86] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior*, vol. 28, no. 2, pp. 117–139, 2004.

[87] R. M. Krauss, P. Morrel-Samuels, and C. Colasante, "Do conversational hand gestures communicate?" *Journal of personality and social psychology*, vol. 61, no. 5, p. 743, 1991.

[88] M. Kipp and J.-C. Martin, “Gesture and emotion: Can basic gestural form features discriminate emotions?” In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, IEEE, 2009, pp. 1–8.

[89] B. de Gelder and N. Hadjikhani, “Non-conscious recognition of emotional body language,” *Neuroreport*, vol. 17, no. 6, pp. 583–586, Apr. 2006.

[90] L. D. Riek, T.-C. Rabinowitch, P. Bremner, A. G. Pipe, M. Fraser, and P. Robinson, “Cooperative gestures: Effective signaling for humanoid robots,” in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, IEEE, 2010, pp. 61–68.

[91] A. Moon, C. A. Parker, E. A. Croft, and H. Van der Loos, “Design and impact of hesitation gestures during human-robot resource conflicts,” *Journal of Human-Robot Interaction*, vol. 2, no. 3, pp. 18–40, 2013.

[92] H. Kozima and H. Yano, “In search of otogenetic prerequisites for embodied social intelligence,” in *Proceedings of the Workshop on Emergence and Development on Embodied Cognition; International Conference on Cognitive Science*, 2001, pp. 30–34.

[93] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, “Affect recognition for interactive companions: Challenges and design in real world scenarios,” *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 89–98, 2010.

[94] M. Scheutz, P. Schermerhorn, and J. Kramer, “The utility of affect expression in natural language interactions in joint human-robot tasks,” in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, ACM, 2006, pp. 226–233.

[95] J. J. Lee, B. Knox, and C. Breazeal, “Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions,” in *2013 AAAI Spring Symposium Series*, 2013.

[96] R. H. Frank, *Passions within reason: The strategic role of the emotions.* WW Norton & Co, 1988.

[97] J. J. Campos, S. Thein, and D. Owen, “A darwinian legacy to understanding human infancy,” *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 110–134, 2003.

[98] N. Frijda, *The emotions*. Cambridge University Press, 1987.

[99] H. A. Simon, "Motivational and emotional controls of cognition.," *Psychological review*, vol. 74, no. 1, p. 29, 1967.

[100] C. Breazeal and R. Brooks, "Robot emotion: A functional perspective," *Who needs emotions*, pp. 271–310, 2005.

[101] J. Velásquez, "Modeling emotion-based decision-making," *Emotional and intelligent: The tangled knot of cognition*, pp. 164–169, 1998.

[102] E. Pot, J. Monceaux, R. Gelin, and B. Maisonnier, "Choregraphe: A graphical tool for humanoid robot programming," in *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, IEEE, 2009, pp. 46–51.

[103] H. C. Traue, F. Ohl, A. Brechmann, F. Schwenker, H. Kessler, K. Limbrecht, H Hoffman, S. Scherer, M. Kotzyba, and A. Scheck, "A framework for emotions and dispositions in man-companion interaction," *Converbal Synchrony in Human-Machine Interaction*, pp. 98–140, 2013.

[104] M. W. Alibali, S. Kita, and A. J. Young, "Gesture and the process of speech production: We think, therefore we gesture," *Language and cognitive processes*, vol. 15, no. 6, pp. 593–613, 2000.

[105] S. Kita, A. Özyürek, S. Allen, A. Brown, R. Furman, and T. Ishizuka, "Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production," *Language and Cognitive Processes*, vol. 22, no. 8, pp. 1212–1236, 2007.

[106] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *International Journal of Social Robotics*, vol. 4, no. 2, pp. 201–217, 2012.

[107] P. N. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral and brain sciences*, vol. 31, no. 5, pp. 559–575, 2008.

[108] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, p. 40, 2012.

[109] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, vol. 5, no. 1_suppl, pp. 123–147, 2001.

[110] X. Downie, C. Laurier, and M. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," in *Proc. 9th Int. Conf. Music Inf. Retrieval*, 2008, pp. 462–467.

[111] X. Hu, V. Sanghvi, B. Vong, P. J. On, C. Leong, and J. Angelica, "Moody: A web-based music mood classification and recommendation system," 2008.

[112] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 448–457, 2008.

[113] T. Hermann, A. Neumann, and S. Zehe, "Head gesture sonification for supporting social interaction," in *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*, 2012, pp. 82–89.

[114] U. Oh, S. K. Kane, and L. Findlater, "Follow that sound: Using sonification and corrective verbal feedback to teach touchscreen gestures," in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '13, Bellevue, Washington: Association for Computing Machinery, 2013, ISBN: 9781450324052.

[115] S. Landry and M. Jeon, "Participatory design research methodologies: A case study in dancer sonification," in *The 23rd International Conference on Auditory Display (ICAD 2017)*, State College, PA, USA, 2017.

[116] N. Schaffert, K. Mattes, and A. O. Effenberg, "A sound design for the purposes of movement optimisation in elite sport (using the example of rowing)," Georgia Institute of Technology, 2009.

[117] M. Schmidmaier, H. Hußmann, and D. M. Runge, "Beep beep: Building trust with sound," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '20, Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–8, ISBN: 9781450368193.

[118] J. L. Alty, D. Rigas, and P. Vickers, "Using music as a communication medium," in *CHI '97 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '97, Atlanta, Georgia: Association for Computing Machinery, 1997, 30–31, ISBN: 0897919262.

[119] J. L. Alty and D. I. Rigas, "Communicating graphical information to blind users using music: The role of context," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '98, Los Angeles, California, USA: ACM Press/Addison-Wesley Publishing Co., 1998, 574–581, ISBN: 0201309874.

[120] C. Bauer and A. Kratschmar, "Designing a music-controlled running application: A sports science and psychological perspective," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '15, Seoul, Republic of Korea: Association for Computing Machinery, 2015, 1379–1384, ISBN: 9781450331463.

[121] N. Warren, M. Jones, S. Jones, and D. Bainbridge, "Navigation via continuously adapted music," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '05, Portland, OR, USA: Association for Computing Machinery, 2005, 1849–1852, ISBN: 1595930027.

[122] L. Hiller, *Music composed with computer [s]: an historical survey*, 18. University of Illinois, 1968.

[123] K. H. Burns, "Algorithmic composition: A definition," *Florida International University*, 1997.

[124] R. Savery, B. Genchel, J. Smith, A. Caulkins, M. Jones, and A. Savery, "Learning from history: Recreating and repurposing sister harriet padberg's computer composed canon and free fugue," *arXiv preprint arXiv:1907.04470*, 2019.

[125] I. Xenakis, "The origins of stochastic music 1," *Tempo*, no. 78, pp. 9–12, 1966.

[126] K. McAlpine, E. Miranda, and S. Hoggar, "Making music with algorithms: A case-study system," *Computer Music Journal*, vol. 23, no. 2, pp. 19–30, 1999.

[127] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep learning techniques for music generation–a survey," *arXiv preprint arXiv:1709.01620*, 2017.

[128] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," *arXiv preprint arXiv:1809.04281*, 2018.

[129] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[130] F. Pachet, "The continuator: Musical interaction with style," *Journal of New Music Research*, vol. 32, no. 3, pp. 333–341, 2003.

[131] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[132] Y. Liu, M. Wang, C. A. Perfetti, B. Brubaker, S. Wu, and B. MacWhinney, “Learning a tonal language by attending to the tone: An in vivo experiment,” *Language Learning*, vol. 61, no. 4, pp. 1119–1141, 2011.

[133] A. E. Thymé-Gobbel and S. E. Hutchins, “On using prosodic cues in automatic language identification,” in *Proc. ICSLP*, vol. 96, 1996, p. 1768.

[134] R. Carlson, A. Friberg, L. Frydén, B. Granström, and J. Sundberg, “Speech and music performance: Parallels and contrasts,” *Contemporary Music Review*, vol. 4, no. 1, pp. 391–404, 1989.

[135] J. Wolfe, “Speech and music, acoustics and coding, and what music might be ‘for’,” in *Proc. 7th International Conference on Music Perception and Cognition*, 2002, pp. 10–13.

[136] I. Chow and S. Brown, “A musical approach to speech melody,” *Frontiers in psychology*, vol. 9, p. 247, 2018.

[137] A. R. Meireles, A. R. Simões, A. C. Ribeiro, and B. R. de Medeiros, “Musical speech: A new methodology for transcribing speech prosody.,” in *Interspeech*, 2017, pp. 334–338.

[138] Y. T. Wang, J. Han, X. Q. Jiang, J. Zou, and H. Zhao, “Study of speech emotion recognition based on prosodic parameters and facial expression features,” in *Applied Mechanics and Materials*, Trans Tech Publ, vol. 241, 2013, pp. 1677–1681.

[139] J. E. Cahn, “The generation of affect in synthesized speech,” *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1990.

[140] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden markov models,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[141] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. M. Meng, and L. Deng, “Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[142] R. Read and T. Belpaeme, “People interpret robotic non-linguistic utterances categorically,” *International Journal of Social Robotics*, vol. 8, no. 1, pp. 31–50, 2016.

[143] M. Tielman, M. Neerincx, J.-J. Meyer, and R. Looije, “Adaptive emotional expression in robot-child interaction,” in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, ACM, 2014, pp. 407–414.

[144] R. Savery, L. Zahray, and G. Weinberg, “Before, between, and after: Enriching robot communication surrounding collaborative creative activities,” *Frontiers in Robotics and AI*, vol. 8, p. 116, 2021.

[145] N. Farris, B. Model, R. Savery, and G. Weinberg, “Musical prosody-driven emotion classification: Interpreting vocalists portrayal of emotions through machine learning,” *18th Sound and Music Computing Conference*, 2021.

[146] R. Savery, L. Zahray, and G. Weinberg, “Emotional musical prosody: Validated vocal dataset for human robot interaction,” in *2020 Joint Conference on AI Music Creativity*, CSMC + MUME, 2020.

[147] V. Sacharin, K. Schlegel, and K. Scherer, “Geneva emotion wheel rating study (report). geneva, switzerland: University of geneva,” *Swiss Center for Affective Sciences*, 2012.

[148] A. K. Coyne, A. Murtagh, and C. McGinn, “Using the geneva emotion wheel to measure perceived affect in human-robot interaction,” in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI ’20, Cambridge, United Kingdom: Association for Computing Machinery, 2020, 491–498, ISBN: 9781450367462.

[149] K. Bischoff, S. Claudiu, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, “Music mood and theme classification - a hybrid approach.,” Jan. 2009, pp. 657–662.

[150] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, “Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2408–2412, 2010.

[151] R. Savery, R. Rose, and G. Weinberg, “Finding shimi’s voice: Fostering human-robot communication with music and a nvidia jetson tx2,” in *Proceedings of the 17th Linux Audio Conference*, 2019, p. 5.

[152] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “Crepe: A convolutional representation for pitch estimation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 161–165.

[153] K. R. Scherer, S. Trznadel, B. Fantini, and J. Sundberg, “Recognizing emotions in the singing voice.,” *Psychomusicology: Music, Mind, and Brain*, vol. 27, no. 4, p. 244, 2017.

[154] K. Sohn, H. Lee, and X. Yan, “Learning structured output representation using deep conditional generative models,” in *Advances in neural information processing systems*, 2015, pp. 3483–3491.

[155] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[156] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, "Midinet: A convolutional generative adversarial network for symbolic-domain music generation," *arXiv preprint arXiv:1703.10847*, 2017.

[157] R. Savery and G. Weinberg, "A survey of robotics and emotion: Classifications and models of emotional interaction," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2020, pp. 986–993.

[158] Z. R. Khavas, S. R. Ahmadzadeh, and P. Robinette, "Modeling trust in human-robot interaction: A survey," in *International Conference on Social Robotics*, Springer, 2020, pp. 529–541.

[159] M. Bretan, G. Hoffman, and G. Weinberg, "Emotionally expressive dynamic physical behaviors in robots," *International Journal of Human-Computer Studies*, vol. 78, pp. 1–16, 2015.

[160] R Savery, L Zahray, and G Weinberg, "Emotional musical prosody for the enhancement of trust in robotic arm communication," in *Trust, Acceptance and Social Cues in Human-Robot Interaction: 29th IEEE International Conference on Robot & Human Interactive Communication*, 2020.

[161] Grand View Research Choice, "Collaborative robots market size, share trends analysis report by payload capacity, by application (assembly, handling, packaging, quality testing), by vertical, by region, and segment forecasts, 2019 - 2025," Grand View Research Choice, Tech. Rep.

[162] M. Tannous, M. Miraglia, F. Inglese, L. Giorgini, F. Ricciardi, R. Pelliccia, M. Milazzo, and C. Stefanini, "Haptic-based touch detection for collaborative robots in welding applications," *Robotics and Computer-Integrated Manufacturing*, vol. 64, p. 101 952, 2020.

[163] E. Rosen, D. Whitney, E. Phillips, G. Chien, J. Tompkin, G. Konidaris, and S. Tellex, "Communicating robot arm motion intent through mixed reality head-mounted displays," in *Robotics research*, Springer, 2020, pp. 301–316.

[164] K. Fischer, "Why collaborative robots must be social (and even emotional) actors," *Techné: Research in Philosophy and Technology*, vol. 23, no. 3, pp. 270–289, 2019.

[165] J. Jost, T. Kirks, S. Chapman, and G. Rinkenauer, "Examining the effects of height, velocity and emotional representation of a social transport robot and human factors

in human-robot collaboration," in *IFIP Conference on Human-Computer Interaction*, Springer, 2019, pp. 517–526.

[166] S. S. Balasuriya, L. Sitbon, M. Brereton, and S. Koplick, "How can social robots spark collaboration and engagement among people with intellectual disability?" In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, ser. OZCHI'19, Fremantle, WA, Australia: Association for Computing Machinery, 2019, 209–220, ISBN: 9781450376969.

[167] L. Desideri, C. Ottaviani, M. Malavasi, R. di Marzio, and P. Bonifacci, "Emotional processes in human-robot interaction during brief cognitive testing," *Computers in Human Behavior*, vol. 90, pp. 331–342, 2019.

[168] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "trust perception scale-hri"," in *Robust Intelligence and Trust in Autonomous Systems*, Springer, 2016, pp. 191–218.

[169] K. Van Dongen and P.-P. Van Maanen, "A framework for explaining reliance on decision aids," *International Journal of Human-Computer Studies*, vol. 71, no. 4, pp. 410–424, 2013.

[170] E. J. de Visser, F. Krueger, P. McKnight, S. Scheid, M. Smith, S. Chalk, and R. Parasuraman, "The world is not enough: Trust in cognitive agents," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Sage Publications Sage CA: Los Angeles, CA, vol. 56, 2012, pp. 263–267.

[171] L. Muralidharan, E. J. de Visser, and R. Parasuraman, "The effects of pitch contour and flanging on trust in speaking cognitive agents," in *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, 2014, pp. 2167–2172.

[172] D. Johnson and K. Grayson, "Cognitive and affective trust in service relationships," *Journal of Business research*, vol. 58, no. 4, pp. 500–507, 2005.

[173] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction*, Elsevier, 2021, pp. 3–25.

[174] R. Savery, R. Rose, and G. Weinberg, "Establishing human-robot trust through music-driven robotic emotion prosody and gesture," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2019, pp. 1–7.

[175] G. Hoffman, "Dumb robots, smart phones: A case study of music listening companionship," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, Paris: IEEE, Sep. 2012, pp. 358–363, ISBN: 978-1-4673-4604-7 978-1-4673-4606-1.

[176] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[177] M. Bretan, G. Hoffman, and G. Weinberg, "Emotionally expressive dynamic physical behaviors in robots," *International Journal of Human-Computer Studies*, vol. 78, pp. 1–16, Jun. 2015.

[178] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, 1980.

[179] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 2018.

[180] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st. O'Reilly Media, Inc., 2009, ISBN: 0596516495, 9780596516499.

[181] D. McNeill, *How language began: Gesture and speech in human evolution*. Cambridge University Press, 2012, pp. 11–12.

[182] B. Sievers, L. Polansky, M. Casey, and T. Wheatley, "Music and movement share a dynamic structure that supports universal expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 110, no. 1, pp. 70–75, Jan. 2013.

[183] P. Toiviainen, G. Luck, and M. R. Thompson, "Embodied Meter: Hierarchical Eigenmodes in Music-Induced Movement," *Music Perception: An Interdisciplinary Journal*, vol. 28, no. 1, pp. 59–70, Sep. 2010.

[184] B. Burger, S. Saarikallio, G. Luck, M. R. Thompson, and P. Toiviainen, "Relationships Between Perceived Emotions in Music and Music-induced Movement," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 5, pp. 517–533, Jun. 2013.

[185] C. Raffel and D. P. W. Ellis, "Intuitive analysis, creation and manipulation of midi data with pretty_midi," in *Proceedings of the 15th International Conference on Music Information Retrieval Late Breaking and Demo Papers*, 2014.

[186] S. M. Mohammad and F. Bravo-Marquez, "WASSA-2017 shared task on emotion intensity," in *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Copenhagen, Denmark, 2017.

[187] A. Lim, T. Ogata, and H. G. Okuno, "Towards expressive musical robots: A cross-modal framework for emotional gesture, voice and music," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, no. 1, p. 3, Jan. 2012.

[188] G. Weinberg, "Expressive digital musical instruments for children," PhD thesis, Massachusetts Institute of Technology, 1999.

[189] A. LaViers, L. Teague, and M. Egerstedt, "Style-based robotic motion in contemporary dance performance," *Controls and Art*, pp. 205–229, 2014.

[190] J. E. Young, M. Xin, and E. Sharlin, "Robot expressionism through cartooning," in *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2007, pp. 309–316.

[191] G. Hoffman and K. Vanunu, "Effects of robotic companionship on music enjoyment and agent perception," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2013, pp. 317–324.

[192] M. A. Boden, "Computer models of creativity," *AI Magazine*, vol. 30, no. 3, pp. 23–23, 2009.

[193] G. Hoffman and G. Weinberg, "Synchronization in human-robot musicianship," in *19th International Symposium in Robot and Human Interactive Communication*, 2010, pp. 718–724.

[194] R. Savery and G. Weinberg, "Shimon the robot film composer and deepscore," *Proceedings of Computer Simulation of Musical Creativity*, p. 5, 2018.

[195] R. Savery, L. Zahray, and G. Weinberg, "Shimon the rapper: A real-time system for human-robot interactive rap battles," *International Conference on Computational Creativity*, 2020.

[196] M. O. Riedl and R. M. Young, "Narrative planning: Balancing plot and character," *Journal of Artificial Intelligence Research*, vol. 39, pp. 217–268, 2010.

[197] D. J. Brooks, C. Lignos, C. Finucane, M. S. Medvedev, I. Perera, V. Raman, H. Kress-Gazit, M. Marcus, and H. A. Yanco, "Make it so: Continuous, flexible natural language interaction with an autonomous robot," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[198] A. Niculescu, B. van Dijk, A. Nijholt, H. Li, and S. L. See, "Making social robots more attractive: The effects of voice pitch, humor and empathy," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 171–191, 2013.

[199] G. Hoffman, J. Forlizzi, S. Ayal, A. Steinfeld, J. Antanitis, G. Hochman, E. Hochendoner, and J. Finkenaur, "Robot presence and human honesty: Experimental evidence," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2015, pp. 181–188.

[200] S. Woods, M. Walters, K. L. Koay, and K. Dautenhahn, "Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach," in *9th IEEE International Workshop on Advanced Motion Control, 2006.*, IEEE, 2006, pp. 750–755.

[201] S. Woods, M. Walters, Kheng Lee Koay, and K. Dautenhahn, "Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach," in *9th IEEE International Workshop on Advanced Motion Control, 2006.*, 2006, pp. 750–755.

[202] M. R. Fraune, S. Kawakami, S. Sabanovic, P. R. S. De Silva, and M. Okada, "Three's company, or a crowd?: The effects of robot number and behavior on hri in japan and the usa.," in *Robotics: Science and systems*, 2015.

[203] K. Walters, D. A. Christakis, and D. R. Wright, "Are mechanical turk worker samples representative of health status and health behaviors in the us?" *PloS one*, vol. 13, no. 6, e0198835, 2018.

[204] S. Tulli, D. A. Ambrossio, A. Najjar, and F. J. R. Lera, "Great expectations & aborted business initiatives: The paradox of social robot between research and industry.," in *BNAIC/BENELEARN*, 2019, pp. 1–10.

[205] K. F. MacDorman and S. O. Entezari, "Individual differences predict sensitivity to the uncanny valley," *Interaction Studies*, vol. 16, no. 2, pp. 141–172, 2015.

[206] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[207] L. A. Pervin, *The science of personality*. Oxford university press, 2003.

[208] R. R. McCrae and P. T. Costa Jr, "Personality trait structure as a human universal.," *American psychologist*, vol. 52, no. 5, p. 509, 1997.

[209] L. M. P. Zillig, S. H. Hemenover, and R. A. Dienstbier, "What do we assess when we assess a big 5 trait? a content analysis of the affective, behavioral, and cognitive processes represented in big 5 personality inventories," *Personality and Social Psychology Bulletin*, vol. 28, no. 6, pp. 847–858, 2002.

[210] J. J. Gross, "Emotion regulation," *Handbook of emotions*, vol. 3, no. 3, pp. 497–513, 2008.

[211] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2008, pp. 2041–2044.

[212] O. P. John, S. Srivastava, *et al.*, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[213] L. Robert, R. Alahmad, C. Esterwood, S. Kim, S. You, and Q. Zhang, "A review of personality in human–robot interactions," *Available at SSRN 3528496*, 2020.

[214] H. Miwa, T. Umetsu, A. Takanishi, and H. Takanobu, "Robot personalization based on the mental dynamics," in *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)(Cat. No. 00CH37113)*, IEEE, vol. 1, 2000, pp. 8–14.

[215] H.-D. Bui, T. L. Q. Dang, and N. Y. Chong, "Robot social emotional development through memory retrieval," in *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*, IEEE, 2019, pp. 46–51.

[216] D. S. Syrdal, K. Dautenhahn, S. N. Woods, M. L. Walters, and K. L. Koay, "Looking good? appearance preferences and robot personality inferences at zero acquaintance.," in *AAAI Spring symposium: multidisciplinary collaboration for socially assistive robotics*, 2007, pp. 86–92.

[217] A.-L. Vollmer, K. J. Rohlfing, B. Wrede, and A. Cangelosi, "Alignment to the actions of a robot," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 241–252, 2015.

[218] D. Conti, E. Commodari, and S. Buono, "Personality factors and acceptability of socially assistive robotics in teachers with and without specialized training for children with disability," *Life Span and Disability*, vol. 20, no. 2, pp. 251–272, 2017.

[219] P. Chevalier, J.-C. Martin, B. Isableu, and A. Tapus, "Impact of personality on the recognition of emotion expressed via human, virtual, and robotic embodiments," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2015, pp. 229–234.

[220] H. S. Ahn, Y. M. Baek, J. H. Na, and J. Y. Choi, "Multi-dimensional emotional engine with personality using intelligent service robot for children," in *2008 International Conference on Control, Automation and Systems*, IEEE, 2008, pp. 2020–2025.

[221] K. Sohn, S. Krishnamoorthy, O. Paul, and M. A. Lewis, "Giving robots a flexible persona: The five factor model of artificial personality in action," in *2012 12th In-*

ternational Conference on Control, Automation and Systems, IEEE, 2012, pp. 133–139.

[222] J.-C. Park, H.-R. Kim, Y.-M. Kim, and D.-S. Kwon, "Robot's individual emotion generation model and action coloring according to the robot's personality," in *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, IEEE, 2009, pp. 257–262.

[223] J. J. Gross, "Emotion regulation: Past, present, future," *Cognition & emotion*, vol. 13, no. 5, pp. 551–573, 1999.

[224] J. J. Gross, G. Sheppes, and H. L. Urry, "Emotion generation and emotion regulation: A distinction we should make (carefully)," *Cognition and emotion (Print)*, vol. 25, no. 5, pp. 765–781, 2011.

[225] A. Gyurak, J. J. Gross, and A. Etkin, "Explicit and implicit emotion regulation: A dual-process framework," *Cognition and emotion*, vol. 25, no. 3, pp. 400–412, 2011.

[226] R. A. Thompson, "Emotion and self," *Socioemotional development*, vol. 36, p. 367, 1990.

[227] P. H. Bucci, X. L. Cang, H. Mah, L. Rodgers, and K. E. MacLean, "Real emotions don't stand still: Toward ecologically viable representation of affective interaction," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019, pp. 1–7.

[228] U. Barańczuk, "The five factor model of personality and emotion regulation: A meta-analysis," *Personality and Individual Differences*, vol. 139, pp. 217–227, 2019.

[229] N. Dael, M. Mortillaro, and K. Scherer, "Emotion expression in body action and posture," *Emotion (Washington, D.C.)*, vol. 12, pp. 1085–101, Nov. 2011.

[230] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, "A very brief measure of the big-five personality domains," *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.

[231] S. Clifford, R. M. Jewell, and P. D. Waggoner, "Are samples drawn from mechanical turk valid for research on political ideology?" *Research & Politics*, vol. 2, no. 4, p. 2 053 168 015 622 072, 2015.

[232] B. Kurdi, S. Lozano, and M. R. Banaji, "Introducing the open affective standardized image set (oasis)," *Behavior research methods*, vol. 49, no. 2, pp. 457–470, 2017.

[233] G. Kochanska, A. E. Friesenborg, L. A. Lange, and M. M. Martel, “Parents’ personality and infants’ temperament as contributors to their emerging relationship.,” *Journal of personality and social psychology*, vol. 86, no. 5, p. 744, 2004.

[234] B.-R. Kim, S.-M. Chow, B. Bray, and D. M. Teti, “Trajectories of mothers’ emotional availability: Relations with infant temperament in predicting attachment security,” *Attachment & human development*, vol. 19, no. 1, pp. 38–57, 2017.

[235] S. Mangelsdorf, M. Gunnar, R. Kestenbaum, S. Lang, and D. Andreas, “Infant proneness-to-distress temperament, maternal personality, and mother-infant attachment: Associations and goodness of fit,” *Child development*, vol. 61, no. 3, pp. 820–831, 1990.

[236] O. P. John, E. M. Donahue, and R. L. Kentle, “Big five inventory,” *Journal of Personality and Social Psychology*, 1991.

[237] M. K. Mount, M. R. Barrick, and J. P. Strauss, “Validity of observer ratings of the big five personality factors.,” *Journal of Applied Psychology*, vol. 79, no. 2, p. 272, 1994.

[238] R. Savery, A. Rogel, and G. Weinberg, “Emotion musical prosody for robotic groups and entitativity,” in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, IEEE, 2021, pp. 440–446.

[239] M. R. Fraune, S. Sherrin, S. Šabanović, and E. R. Smith, “Is human-robot interaction more competitive between groups than between individuals?” In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2019, pp. 104–113.

[240] M. A. Hogg, D. K. Sherman, J. Dierselhuis, A. T. Maitner, and G. Moffitt, “Uncertainty, entitativity, and group identification,” *Journal of experimental social psychology*, vol. 43, no. 1, pp. 135–142, 2007.

[241] A. L. Blanchard, L. E. Caudill, and L. S. Walker, “Developing an entitativity measure and distinguishing it from antecedents and outcomes within online and face-to-face groups,” *Group Processes & Intergroup Relations*, vol. 23, no. 1, pp. 91–108, 2020.

[242] N. Dasgupta, M. R. Banaji, and R. P. Abelson, “Group entitativity and group perception: Associations between physical features and psychological judgment.,” *Journal of personality and social psychology*, vol. 77, no. 5, p. 991, 1999.

[243] A. Bera, T. Randhavane, E. Kubin, A. Wang, K. Gray, and D. Manocha, “The socially invisible robot navigation in the social world using robot entitativity,”

in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 4468–4475.

[244] E. Castano, V. Yzerbyt, and D. Bourguignon, “We are one and i like it: The impact of ingroup entitativity on ingroup identification,” *European journal of social psychology*, vol. 33, no. 6, pp. 735–754, 2003.

[245] E. Castano, “On the advantages of reifying the ingroup,” *The psychology of group perception: Perceived variability, entitativity, and essentialism*, pp. 381–400, 2004.

[246] D. L. Hamilton, S. J. Sherman, and L. Castelli, “A group by any other name—the role of entitativity in group perception,” *European review of social psychology*, vol. 12, no. 1, pp. 139–166, 2002.

[247] M. R. Fraune, S. Sherrin, S. Sabanović, and E. R. Smith, “Rabble of robots effects: Number and type of robots modulates attitudes, emotions, and stereotypes,” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 109–116.

[248] A. M. Abrams and A. M. Rosenthal-von der Pütten, “I–c–e framework: Concepts for group dynamics research in human-robot interaction,” *International Journal of Social Robotics*, pp. 1–17, 2020.

[249] M. Chita-Tegmark, T. Law, N. Rabb, and M. Scheutz, “Can you trust your trust measure?” In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI ’21, Boulder, CO, USA: Association for Computing Machinery, 2021, 92–100, ISBN: 9781450382892.

[250] K. Kurebayashi, L. Hoffman, C. S. Ryan, and A. Murayama, “Japanese and american perceptions of group entitativity and autonomy: A multilevel analysis,” *Journal of Cross-Cultural Psychology*, vol. 43, no. 2, pp. 349–364, 2012.

[251] B. Lickel, D. L. Hamilton, G. Wieczorkowska, A. Lewis, S. J. Sherman, and A. N. Uhles, “Varieties of groups and the perception of group entitativity.,” *Journal of personality and social psychology*, vol. 78, no. 2, p. 223, 2000.

[252] D. J. Hauser and N. Schwarz, “Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants,” *Behavior research methods*, vol. 48, no. 1, pp. 400–407, 2016.

[253] B. Irfan, A. Ramachandran, S. Spaulding, D. F. Glas, I. Leite, and K. L. Koay, “Personalization in long-term human-robot interaction,” in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 685–686.