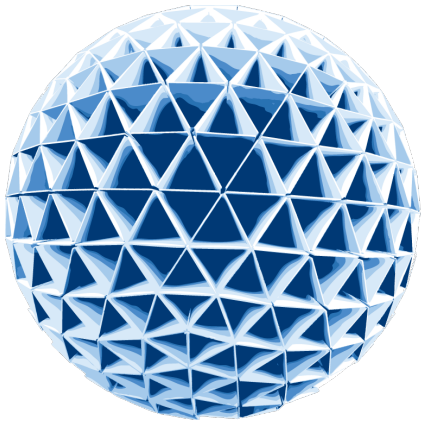
Exploring Roboticness and Applicability of Voices to Social Robots

Bachelor's Thesis

BSc. Technology & Liberal Arts and Sciences

by

Pola Z. Labedzka



Supervisors

Daily Supervisor: **dr. Khiet P. Truong**

UCT Supervisor: **dr. Marcus Gerhold**

Dragos. A. Balan and **Hideki Garcia Goo**

June 2024

Enschede, the Netherlands

University of Twente, University College Twente

Abstract

Robots, especially social robots, are becoming increasingly popular in today's world. With that, the need to understand the perception of robotic speech and its appropriate design arises. This project aims to understand and model what makes a voice perceived as 'robotic' as well as what makes it suitable for social robots. This research has been split into two studies. In the first one, people's impressions of *roboticess* and the suitability of voices to social robots, along with associations with robots, have been gathered in a form of an online survey. In the second one, an attempt was made to quantify and measure those impressions using speaker embeddings, fixed-dimensional vectors that capture the speaker's identity, and vocal parameters. The results support the earlier findings that the less robotic a voice is, the more applicable it becomes to a social robot. Moreover, Study I also revealed that people still often associate robots with those from the media. This suggests that media is a useful source of knowledge for understanding what makes a voice sound robotic. The findings of Study II further support this conclusion, as it was found that vocal parameters related to fundamental frequency and loudness are the ones primarily related to the *roboticness* of a given voice. On the other hand, thematic analysis of perceived factors contributing to *roboticness*, gathered in Study I, revealed less measurable descriptors of voices with such quality. Henceforth, in Study II, speaker embeddings were employed to explore whether they encode *roboticness*. The results of this study indicated that speaker embeddings contained information related to the perception of robotic voices to some extent. However, this finding needs validation through further research, which should address significant limitations not covered in this study.

Acknowledgements

Firstly, I would like to thank Khiet Truong for her unwavering patience, support, and invaluable guidance over the course of this project. Thank you also to Marcus Gerhold for his steadfast support and belief in my abilities. Your confidence in me has been a constant source of motivation. I would also like to thank Hideki and Dragos for always being willing to answer my questions, bringing me new ideas, and showing such a strong investment in my work. Your enthusiasm and expertise have been greatly appreciated. Moreover, I would like to express my sincere thanks to Aki Kunikoshi, whose inspiration for this project and significant support in helping me wrap my head around complex concepts have been indispensable. Lastly, I would like to extend my heartfelt appreciation to all my friends and the UCT community for consistently standing by my side and having faith in me.

Contents

1 Introduction **1**

2 Background **3**
2.1 'Roboticness' . . . 3
2.2 Robotic Speech in Media . . . 3
2.3 Social Robots and Their Voice Design . . . 5
2.4 Voice Analysis . . . 7
2.4.1 Vocal Parameters . . . 7
2.4.2 Timbre . . . 7
2.4.3 Speaker Embeddings . . . 8

3 Present Study **9**

4 Study I: Perception of Robotic Voices **11**
4.1 Methodology . . . 11
4.1.1 Participants . . . 12
4.1.2 Materials . . . 13
4.1.3 Measures . . . 17
4.1.4 Setup and Procedure . . . 17
4.1.5 Data Analysis . . . 19
4.2 Results . . . 19
4.2.1 Summary of the Ratings of *Roboticness* and Suitability . . . 19
4.2.2 Relationship between 'Roboticness' and 'Suitability' . . . 21
4.2.3 Factors that Influence Perceived *Roboticness* . . . 22
4.2.4 Associations with Robots . . . 23

5 Study II: Speaker Embeddings and Vocal Parameters Analysis **24**
5.1 Methodology . . . 24
5.1.1 Speaker Embeddings . . . 24
5.1.2 Acoustic Parameters . . . 25
5.2 Analysis . . . 25
5.2.1 Speaker Embeddings . . . 25
5.2.2 Acoustic Parameters . . . 26
5.3 Results . . . 26
5.3.1 t-SNE Vizualization . . . 26
5.3.2 Predictions Based on Speaker Embeddings . . . 28
5.3.3 eGeMAPS Features: Simple Linear Regression . . . 29
5.3.4 eGeMAPS Features: Recursive Feature Elimination . . . 29
5.4 Discussions and Limitations . . . 29
5.4.1 Relationship Between *Roboticness* and Suitability to Social Robots 30

5.4.2 People's Perception of *Roboticness* 30
5.4.3 Associations with Robots 31
5.4.4 Encoding of the Perception of Robots in Speaker Embeddings . . . 32
5.4.5 Analysis of Perceptions about Robotic Voices Through eGeMAPS Features 34

6 Conclusion **35**

7 Future Research **36**

8 Contextual Exploration **36**
8.1 Interdisciplinarity 36
8.2 Small Scale Implications 37
8.3 Large Scale Implications 37

A eGeMAPS Features Analysis **46**
A.1 RFE Statistics 51
A.2 35 Best Features Selected by RFE 54

1 Introduction

There is no doubt that robots become more present in today's world. They are no longer just figures in science fiction movies, but are functional entities that are present in the real world.

"A robot is an autonomous machine capable of sensing its environment, carrying out computations to make decisions, and performing actions in the real world." – Guizzo (2018)

"(especially in science fiction) a machine resembling a human being and able to replicate certain human movements and functions automatically." – (Oxford University Press, 2005, p. 302)

"A machine controlled by a computer that is used to perform jobs automatically" – (Cambridge University Press, n.d.)

From the definitions above, it is evident that the main feature that characterizes a robot is that it performs tasks automatically. Additionally, some sources expand on this definition by emphasizing that robots are designed to support or replicate specific human capabilities. Furthermore, robots can be categorized into different types. Guizzo (2018) lists eighteen categories of robots. Some of them are ones that the general public might have had experience with, such as consumer robots, with a popular example of a *Roomba*, the vacuuming robot, and some with more niche applications, such as an aquatic robot. One can imagine that each of these robot types also differs in sound design – robots that work in factories do not need as extensive sound design as robots that are designed to interact with humans in social settings and environments. Those latter ones, also called 'social robots' are becoming more advanced as the technology develops (Duffy et al., 1999; Naneva et al., 2020). Since such robots are designed to interact with people who are not necessarily experts in the field of robotics, identifying ways in which their acceptance could be improved is crucial for the proper development of such technology (Naneva et al., 2020; Sheridan, 2020). Research has shown that focusing on visual appearance as well as speech characteristics can significantly improve the overall acceptance of such robots (McGinn & Torre, 2019). Furthermore, there is an ongoing debate regarding the level of anthropomorphism social robots should achieve (Li & Suh, 2021). Even though a few studies suggest a preference for highly human-like voices in social robots, the topic appears to be underexplored (Li & Suh, 2021; Sheridan, 2020; Wilson & Moore, 2017).

Therefore, identifying measurable vocal features that make voices sound robotic or make them more applicable to social robots could be useful to understand the human perception of robotic speech better and hence improve the design of robots. Nevertheless, some properties of speech, such as the timbre that allows people to distinguish the sound of two speakers or instruments from each other are not precisely measurable with the use of vocal

parameters. Recent advancements in speech technology offer promising avenues for further enhancing the analysis of speech. One example is the technology of speaker embeddings, which represent the speaker's identity and voice's unique characteristics in the form of fixed-size vectors (Jakubec et al., 2024). By better understanding what information about robotic speech is captured in speaker embeddings, they could potentially be used for modelling and designing robotic voices in the future.

This project has been divided into two studies whose aim was to understand the human perception of robotic speech and different voices' applicability to social robots. **Study I** focused on gathering people's impressions of robots and different sounds in the form of a listening test and analysing them. The goal of **Study II** was to quantify and further understand those impressions by applying complex feature extraction methods.

Study I was based on answering the following research question: **RQ 1**: *How do people perceive robotic voices?* Primarily, this study examined the level of anthropomorphism that the voices of social robots should attain by addressing the first formulated subquestion– **RQ 1.1**: *To what degree is the suitability of a voice for a social robot dependent on the roboticness of the voice?*

Henceforth, this research has focused on the properties of robotic speech. Firstly, it was argued that there exists a certain *roboticness* characteristic of a voice that people can identify. Guided by the second subquestion, **RQ 1.2**: *What perceived factors influence individuals' assessments of roboticness of voices?*, it was investigated what perceived aspects of a given voice make it sound more robotic than others.

Furthermore, in order to define whether media can be a source of knowledge about robotic speech, the third subquestion was posed – **RQ 1.3**: *What are people's associations with robots?* This question explored what kind of mental models people have of robots, helping to interpret the results coming from answering the other subquestions.

Subsequently, in Study II, it was hypothesized that the *roboticness* is a property of voice that can be quantified and measured. Similarly, it was hypothesized that the suitability of a voice to a social robot stems from its certain vocal properties and can also be quantified and measured. Firstly, it was investigated whether speaker embeddings, multidimensional vectors that represent a speaker's identity, encode information about *roboticness* and suitability of a voice to a social robot. This was guided by the following research question – **RQ 2**: *To what extent is information about roboticness and the applicability of voices to social robots captured in speaker embeddings?* and two subquestions – **RQ 2.1**: *To what extent is information about roboticness captured in speaker embeddings?* and **RQ 2.2**: *To what extent is information about perceived suitability to a social robot captured in speaker embeddings?*

Furthermore, vocal features of the sounds that contribute to a certain voice sounding more robotic and more suitable to a social robot were investigated. This was based on

the final research question – **RQ 3**: *What vocal parameters influence the perception of roboticness and the applicability of voices to social robots?* and its two subquestions: **RQ 3.1**: *What vocal parameters influence the perception of roboticness?* and **RQ 3.2**: *What vocal parameters influence the perceived suitability of a voice for a social robot?*

2 Background

2.1 'Roboticness'

While it is apparent that speech can vary from sounding entirely natural and human-like to distinctly robotic, identifying the precise characteristics that contribute to its *roboticness* remains somewhat challenging. Literature on human-robot interaction often refers to certain voices as *natural, synthesized, artificial, or robotic* (Ehret et al., 2021; Gessinger et al., 2022; Schreibelmayr & Mara, 2022). Some papers mention the speech characteristics of artificial voices as *monotonous* or *lacking emotion* but little investigation of what exact attributes and prosodic features make certain speech sound robotic can be found in the literature. Kühne et al. (2020) presented participants with two synthetic voices and one human one. Among other questions, they asked participants to rate the sounds on *human-likeness* and provide explanations for their ratings. Conducting qualitative analysis on the responses, researchers have found that *intonation, sound, emotion, and imageability/embodiment* were the groups of factors that allow to differentiate human speech from a synthetic one. An interesting finding that can also shed some light on what causes the *roboticness* of speech comes from a study by Ehret et al. (2021) where participants rated male synthetic voices as more natural sounding as compared to female synthetic ones. Moreover, the main finding of their study was that inadequate prosody, specifically related to the accent placement, decreased the *naturalness* of speech indicating that not only the overall pitch frequencies but also the accent or intonation of pitch affect how human-like certain voice is perceived. The intonation of pitch was also found to play a role in the study by Bakardzhiev (2022), where the naturalness of a voice assistant's speech in relation to prosodic differences, specifically to prosodic fluctuations, was explored. It was found that a voice with a flat pitch was significantly less naturalistic.

2.2 Robotic Speech in Media

As robots originate from science fiction, many robots can be found in the media, especially in movies. Some of the most famous examples include R2-D2 and C-3PO (robots from the film series *Star Wars*) and WALL-E (Kriz et al., 2010). Such robots are a part of today's culture, and some studies indicate that science fiction representations and interfaces of robots shape individuals' expectations and understanding of real-life interactions and the design of robots (Kriz et al., 2010; Savela et al., 2021). Nevertheless, it must be acknowledged that, as with any quickly developing technology, associations with robots

might change depending on the level of experience that people have with them. As explored by Oliveira and Yadollahi (2024), robots are growing in popularity, but the vast majority of the more advanced models with interaction capabilities are still primarily used for human-robot interaction research.

Therefore, under the assumption that people's associations with robots are strictly connected to media, understanding the aforementioned *roboticness* of speech, it can be beneficial to explore how robotic speech is created in movies. Based on work by Rose (2012), Wilson and Moore (2017) list three main ways in which fictional voices can be produced: *a)* employing a skilled actor who is able to talk in a desired voice, *b)* artificially synthesizing a voice with distinct attributes, *c)* adjusting a voice in post-production. The last method can comprise many different techniques.

Wilson and Moore (2017) list sixteen of those techniques that differ in their exact methodology and can be applied together to create remarkably different voices. Wilson and Moore (2017) further analyse voices of 93 fictional characters from different movies. Among those, they distinguish between robots, aliens, and cartoon characters and analyse differences between their voices and some control (human) voices. Their findings suggest that both alien and robotic voices have distinctly different voice qualities, measured as higher pitch shimmer and larger ranges of mean pitch as compared to human voices. Moreover, character voices overall significantly differed in voice breaks per second from the control voices. Furthermore, Wilson and Moore (2017), found that voices were better indicators of characters' personalities than their appearance. Comparatively, Latupeirissa et al. (2019) who analysed robotic voices from five movies, found a strong relationship between the physical appearance of robots and their sonic presence. Moreover, they used the Long Time Average Spectrum (LTAS) to differentiate between the characteristics of robotic and human voices. The results revealed that robotic voices had a broader frequency spectrum than humans and that their fundamental frequency (f0) was either higher or lower than that of humans. This finding is also strictly related to the one of Wilson and Moore (2017) since the pitch is defined as human's perception of f0 (Bäckström et al., 2022).

When investigating what exact vocal features are adapted in post-production to make a human voice sound like a robot, Wilson and Moore (2017) point at the existing filters available online. Platforms that are popular and available for use include VoiceWave (n.d.), Voicemod (2019), *Voxal* (NCH Software, n.d.), and many others (Lee, 2024). Nevertheless, most of the platforms do not share the details of how the voices are being adapted. One exception is the Voxal platform, on which the pitch is shifted 0.9, a 20-ms echo (delay) is added and the sound is amplified by 64% (see Figure 1).

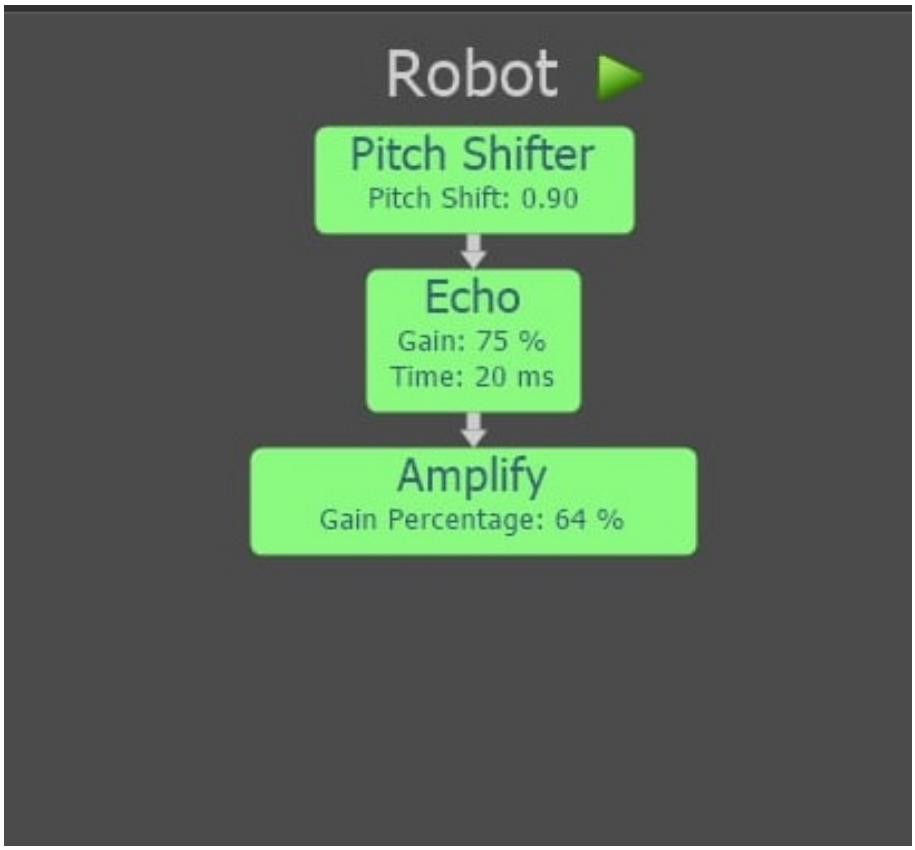


Figure 1: Voxal NCH: a screenshot of a 'Robot' filter

The list below summarizes all the characteristics of voice that were found to make it sound robotic from the sources described above:

- voice breaks per second
- pitch shimmer
- mean pitch range
- f0 higher or lower than humans
- broader frequency spectrum
- delay

2.3 Social Robots and Their Voice Design

Social robots are becoming more popular in today's world due to the development of technology. Duffy et al. (1999) define social robots as physical entities that can behave and interact within complex social environments, benefiting both themselves and the community. Naneva et al. (2020) take on a similar definition, emphasizing the fact that social robots need to have features that allow humans to perceive them as social entities

and are able to interact with humans via a social interface using verbal or non-verbal cues.

Social robots display a broad spectrum of features and can be categorized based on various criteria, for example, on their application. The literature primarily analyses social robots' applications in healthcare, household, tourism services, and education (Fosso et al., 2023). Furthermore, social robots also differ in the level of human resemblance ranging from *humanoid* robots such as AMECA (Arts, 2023) to robots that resemble animals such as AIBO (Moon, 2001).

With robots varying across so many dimensions, differences in the requirements for their voice and sound design can naturally be anticipated. Research has proven that the exact characteristics of voice for a social robot largely depend on appearance or the setting in which the robot is designed to interact. For example, Dou et al. (2020) found that adult female and male voices were more likely to be accepted for an educational robot whereas adult male and child voices were preferred for shopping reception and domestic robots. Moreover, research by McGinn and Torre (2019) demonstrated that both the gender and naturalness of a voice influenced participants' choices when matching voices to pictures of robots with different appearances. Furthermore, Niculescu et al. (2013) found that a robot with a higher pitch was generally rated better in terms of voice attractiveness, general appearance and personality. However, those results might have been influenced by the fact that participants engaged with a physical female robot.

On the other hand, social robots share a common ability to interact with humans, and since verbal cues are fundamental to human interaction, it is logical to anticipate shared principles in speech design among such robots. Nevertheless, little research has focused on generalizing the characteristics that would make a voice applicable to social robots across various contexts.

Voices of social robots can also be analysed in terms of the idea of *roboticness* which was described in the Section 2.1. As humanoid social robots, i.e., those resembling humans in visual terms, become more prevalent, the question arises as to to what extent their voices should resemble those of humans. Studies that focus on exploring human perception of robots often refer to the *uncanny valley* effect that has been proposed by Mori (1970). It reflects that people's preference towards artificial objects does not necessarily increase as they become more human-resembling but rather there is a point at which very realistic, yet containing some evident artificial factors, objects start causing a feeling of eeriness or creepiness (Mara et al., 2022; Mori, 1970). Multiple studies can be found that validate the *uncanny valley* effect for the visual characteristics of social robots (Mende et al., 2019; Tung, 2016). Nevertheless, most literature on the voice characteristics of robots seems to conclude that humans generally favour human speech in the context of artificial agents. Gurung et al. (2023) have presented participants with videos of an artist – one human and two artificial ones with varying levels of naturalness – and found that overall human speech

was viewed more positively than machine-like voice and that inclusion of the artificial voices negatively impacted the perception of the virtual agents. Furthermore, human voices were also clearly preferred by the participants of the aforementioned study by Kühne et al. (2020). Additionally, the more human-like a voice was, the less eerie it was perceived, showing that the *uncanny valley* hypothesis did not hold for speech. Similar results were also obtained by Schreibelmayr and Mara (2022) who found a positive relationship between voices resembling human ones and user acceptance of that voice being used by robots in all considered applications, e.g., for companionship or entertainment.

2.4 Voice Analysis

The aim of voice analysis is to identify certain characteristics of speech that are beyond the linguistic content (Farnsworth, 2023). This entails dividing the sound into segments and extracting different features, such as prosody and intonation, which are then utilized to explain higher-level speech characteristics such as the emotional state of a speaker or, as in the case of this research, *roboticness* and suitability to social robots.

2.4.1 Vocal Parameters

Voice analysis can be done by extracting certain vocal features (Farnsworth, 2023). However, they are often extracted in different combinations and using different techniques (Eyben et al., 2016). Eyben et al. (2016) attempt to mitigate that issue by constructing the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) that forms a toolkit to extract vocal features in a standardized way. The toolkit consists of 18 'Low-Level Descriptors' (LLD) of voice from three different groups of parameters (i) frequency-related, (ii) energy/amplitude related, and (iii) spectral (balance) ones.

2.4.2 Timbre

Vocal analysis can also be looked at from the perspective of linguistic, paralinguistic and extralinguistic 'behaviours' defined by Laver (1994). According to Lu et al. (2023), linguistic behaviour is related to the content of the spoken text, and paralinguistic behaviour explains the speaker's current state, e.g. emotional state, and can be measured in terms of prosody. Extralinguistic behaviour, however, defines the speaker's vocal identity or fingerprint and is expressed by vocal timbre. According to the American National Standards Institute (1973), timbre reflects the quality of auditory sensation that makes people perceive two sounds as different even when presented with equal loudness and pitch. Timbre is of particular interest to this project, as it can provide crucial insights into why certain voices are distinguishable as more robotic from others.

Research into timbre extraction is multidisciplinary and complex (Siedenburg et al., 2019). Nevertheless, it is a common practice to define timbre just by mel-frequency cepstral

coefficients (MFCC) which are also part of the GeMAPS parameters(Eyben et al., 2016; Hansen & Hasan, 2015; Siedenburg et al., 2019). In contrast, speaker embeddings, a more advanced deep learning approach, offer a powerful way to capture speaker identity. Therefore, adhering to the definition by the American National Standards Institute (1973), speaker embeddings, in a way, also encode the timbre of a given voice.

2.4.3 Speaker Embeddings

Speaker embeddings, fixed-size vectors that represent the speaker's identity, have the aim of capturing the speaker's speech characteristics and therefore are often utilized for speaker verification and speaker recognition tasks (Jakubec et al., 2024; Lüscher et al., 2023; Stan, 2022).

Newer research papers explore applications of speaker embeddings to voice conversion, i.e. technology that, given a target speaker's sound sample, can convert a source speaker's speech to sound like the target speaker while preserving the semantic content of the source speaker's speech sample (Lin et al., 2023). Jia et al. (2019) concatenate the output of the speaker embeddings (*d-vectors*) with a TTS synthesizer, allowing for a transfer of the speaker style. Similarily, Lin et al. (2023) make use of the pre-trained `Wav2vec 2.0` model to decouple the semantic characteristics of the source speech and the `WavLM` model to extract the target speaker representation. They later combine these semantic and speaker features, utilizing the `FastSpeech2` model to synthesize speech with the semantic content of the source speaker but the timbre of the target speaker. Shaheen et al. (2023), however, take advantage of speaker embeddings' ability to encode emotions. They identify the components of the vectors associated with emotions and prosody and use that information to synthesize emotional speech.

Because of this utilization of speaker embeddings in text-to-Speech (TTS) applications, they are an interesting topic to explore in the realm of robotic TTS, since it is a technology widely employed for generating robotic speech (Su et al., 2023). One example of employing speaker embeddings for robotic TTS is the attempt to generate a genderless voice for a robot by Yu et al. (2022). By understanding better what information of robotic speech is captured in speaker embeddings, they could potentially be used for modelling and designing robotic voices in the future.

The literature presented above highlights some research gaps to be addressed in this project. Firstly, there seems to be a lack of consensus regarding what makes a certain voice sound robotic. Research suggests that media and robotic filters could potentially provide insights into human understanding of robotic speech but it has to be first established whether such voices remain a source of knowledge for human-robot interaction in this context. Furthermore, there appears to be a lack of objective, measurable methods

to assess this *roboticness*. Secondly, the literature highlights the need for research on the voices of social robots – the level of their anthropomorphism and similarly as with *roboticness* quantifiable measures to assess and understand the suitability of voices to social robots.

3 Present Study

As mentioned in the Introduction, this project was divided into two studies. The first study gathered people's impressions of robotic voices, while the second study analyzed these voices by extracting features and relating them to the impressions gathered in the first study. In the first study, Study I, the following research question was addressed:

Research Question 1: How do people perceive robotic voices?

This research question was investigated by addressing a series of sub-questions. Firstly, the expected level of anthropomorphism of the voice of a social robot was investigated with the RQ 1.1:

RQ 1.1: To what degree is the suitability of a voice for a social robot dependent on the roboticness of the voice?

It was hypothesized that while human-like voices would generally be preferred for social robots, following the conclusions from the literature outlined in the Section 2.3 (Gurung et al., 2023; Kühne et al., 2020; Schreibelmayr & Mara, 2022). Nevertheless, it was kept in mind that voices sounding entirely human-like could not always lead to the highest suitability ratings, as people would not necessarily want robots to mimic humans (Mende et al., 2019; Tung, 2016).

Subsequently, the second subquestion aimed to delve into human perception of the concept of *roboticness* in voices. Specifically, it explored how people perceive this concept in different voices:

RQ 1.2: What perceived factors influence individuals' assessments of *roboticness* of voices?

The hypothesized descriptors individuals might have referred to when assessing voice pleasantness were related to *intonation, sound, emotion, and imageability/embodiment* themes identified in a study by Kühne et al. (2020). It was especially expected to see descriptors associated with pitch fluctuations and monotonicity of a sound following other sources of literature outlined in the Section 2.1. The *emotion* factor was assumed not to be evident in the analysis as the participants were specifically asked not to pay attention to the emotions expressed.

As mentioned before, one of the goals of this project was to measure and quantify the

roboticness of voices. Moreover, the literature suggests that sources of robotic voices from media or robotic filters could potentially offer insights into understanding what exact features make a robotic voice be perceived as such (Latupeirissa et al., 2019; Wilson & Moore, 2017). Nevertheless, with more robots being embedded in society and artificial agents becoming part of reality, the associations with robots and their voices might be changing. In order to evaluate whether the concept of *roboticness* can be understood through the lens of robots from media, the third subquestion was posed:

RQ 1.3: What are people's associations with robots?

It was hypothesized that the associations might come from either media, in particular movies, or from personal interactions and experiences with robots, depending on the level of familiarity with robots (Kriz et al., 2010; Oliveira & Yadollahi, 2024; Savela et al., 2021).

Study II, utilizing the impressions of voices gathered in Study I, focused on quantifying and analysing both *roboticness* and suitability to social robots, utilizing two methods of feature extraction. The second research question of this project focused on assessing the applicability of the aforementioned speaker embeddings for understanding the *roboticness* and the applicability of voices to social robots:

Research Question 2: To what extent is information *roboticness* and the applicability of voices to social robots captured in speaker embeddings?

This research question was further split into the following subquestions focusing on *roboticness* and the suitability of voices to social robots, respectively:

RQ 2.1: To what extent is information about *roboticness* captured in speaker embeddings?

RQ 2.2: To what extent is information about perceived suitability to a social robot captured in speaker embeddings?

Since, to the researcher's knowledge, there are no existing studies that have previously investigated the robotic properties of voices through speaker embeddings, an explicit hypothesis has not been formed. Nevertheless, as speaker embeddings capture speaker-specific, timbral characteristics, it was thought that they could be used to identify voices that are commonly recognized as robotic. Moreover, speaker embeddings were found to often encode information about features that could be relevant for the levels of *roboticness* or applicability to social robots, such as recording conditions or speaking style of the utterances (Stan, 2022). Hence, it was reasonable to assume that speaker embeddings could capture the levels of *roboticness* and applicability to social robots to some extent.

The final, third, research question was aimed at identifying concrete, measurable vocal characteristics of speech that would make it sound more robotic or more suitable for social robots:

Research Question 3: What vocal parameters influence the perception of *roboticness* and the applicability of voices to social robots?

Similarly to RQ 2, this research question was divided into two separate subquestions, one concerning the roboticness of the voice and the other regarding its suitability for a social robot:

RQ 3.1: What vocal parameters influence the perception of *roboticness*?

RQ 3.2: What vocal parameters influence the perceived suitability of a voice for a social robot?

First of all, it was expected that the parameters contributing to a voice sounding more robotic would negatively impact the perceived applicability of it to a social robot, which stemmed from the hypothesized negative relationship between *roboticness* and suitability (Gurung et al., 2023; Schreibelmayr & Mara, 2022; Wilson & Moore, 2017). Furthermore, based on the characteristics identified in the Section 2.2, it was expected for frequency-related features, especially the fundamental frequency (f0) to contribute to the *roboticness* of the sound (Kühne et al., 2020; Latupeirissa et al., 2019). Moreover, it was also anticipated that mel-frequency cepstral coefficients (MFCC) would differ between robotic and human-sounding voices since they encode the timbre of the voice (Eyben et al., 2016). And, as mentioned before, it was believed that it is the timbre of a voice that allowed people to distinguish between human and robotic speech.

4 Study I: Perception of Robotic Voices

4.1 Methodology

Firstly, a corpus of sound samples was constructed by selecting sounds from existing databases of either robotic or human speech. Subsequently, an online listening test was constructed and distributed to participants. Its aim was to collect people's impressions of the voices by gathering ratings of the *roboticness* and suitability to a social robot of each of the preselected sound samples.

Prior to conducting the online survey, the research had been approved by the Ethics Committee on Computer & Information Science (EC-CIS) at the University of Twente.

4.1.1 Participants

65 valid and complete responses to the survey have been recorded. Among them ∼ 56% (36) reported identifying as female, ∼ 42% (27) as male, 1 person reported not identifying with any particular gender, and 1 person wished not to disclose their gender identity. ∼ 54% of the participants were in the age group between 16 and 24, followed by ∼ 18% in the age group 45 to 54. Participants identified with cultural values, norms, and practices from 17 different countries with the most common being Poland (∼ 26%), The Netherlands (∼ 23%), and Latvia (∼ 17%).

As visible in the figure below, above half of the respondents reported having at least some level of familiarity with robots and only 3.1% (2 respondents) reported being extremely unfamiliar with them. When asked about the source of their familiarity with robots, most respondents mentioned educational-related activities such as university or high school. Additionally, 4 respondents reported working with robots, and another 4 mentioned being involved in robotics-related projects. Moreover, 7 respondents mentioned social media or a specific type of social media, while 3 referred to media in general and 2 to books, movies, and TV games. 2 respondents pointed to the internet in general as their source of knowledge about robots, and 5 people mentioned their familiarity coming from the news. Furthermore, 7 respondents reported interacting with artificial agents in their daily lives – 4 with chatbots or voice assistants, 2 with cleaning robots, and 1 with a robot in a supermarket.

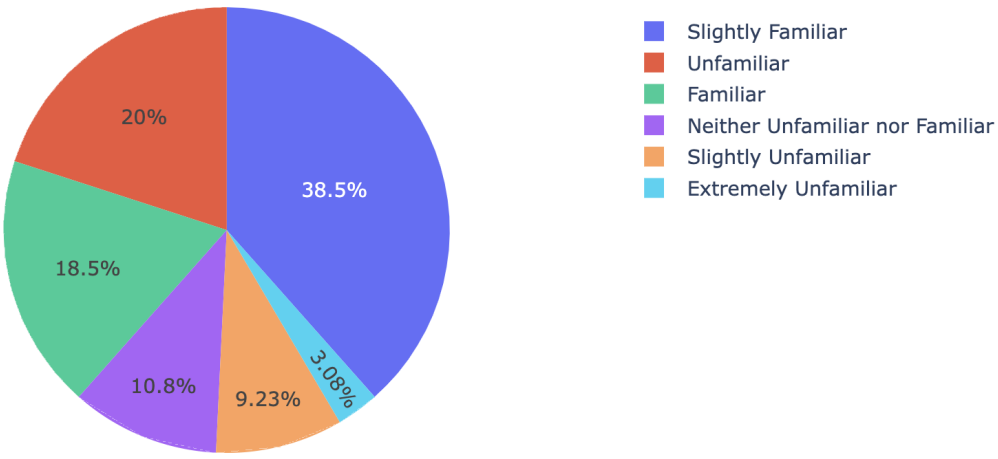


Figure 2: Reported level of familiarity with robots

4.1.2 Materials

To conduct the listening test, an appropriate set of voices had to be constructed. When selecting voices for the created corpus, several principles were followed. Foremost, the voices selected ranged from human-like, i.e., from corpora with human speech, to very robotic-like, i.e., from corpora with synthetic sounds, to ensure a diverse range of voice characteristics and potentially be able to identify factors that make certain voices sound robotic. Second, it was ensured to include robotic voices from different platforms and produced in various ways to represent a wide landscape of robotic speech. Moreover, the decision to focus on semantic-free utterances was made. It was believed that diverse semantic content would affect people's perceptions and beliefs about a certain persona. Given the limited number of corpora with robotic speech available, it was unfeasible to produce the same linguistic content while ensuring the properties outlined before within the scope of this project. This naturally causes considerable limitations, especially regarding the speaker embedding extraction method, which is discussed in the Discussions and Limitations section. Finally, it was decided to include no more than 500 sounds in the final set of sounds included in this research due to the scope and time sensitivity of this project.

The next paragraphs will explain in detail the procedure of choosing sounds from each corpus included in this research. The final summary of the sound samples included is presented in Table 1.

Montreal Affective Voices

Human voices employed mostly came from the Montreal Affective Voices set (hereinafter referred to as MONTREAL), which includes 90 emotional, nonverbal sounds produced by 10 actors (Belin et al., 2008). From the 90 sounds, 8 sounds were excluded since the speaker embeddings extraction method failed to extract the representation from them (with the possible reason being that the recordings were too short, i.e., $< 0.3ms$), which resulted in 82 sound samples from this corpus being included in the listening test.

Subsequently, to introduce speech that was presumed to sound more robotic, samples from the Montreal Affective Voices were modified using code adapted from the PythonAudioEffects GitHub repository (Nxbyte, 2024). The modification included a pitch shift of 0.9, a delay of 0.02 milliseconds and increasing volume 4 times based on the robotic effect from NCH Software (n.d.). This resulted in the creation of another set of sounds, hereinafter referred to as Montreal Affective Voices Robotic or MONTREAL ROB. The selected sound samples were included to assess whether the robotic filters make human speech sound robotic and whether people can identify the human speech behind the filter and rate the sounds as more human-like compared to entirely synthetic voices.

Willow Garage HRI Sound Library

Next, the Willow Garage HRI Sound Library (hereinafter referred to as WILLOW) made publicly available on GitHub was included due to diverse robotic sound samples (Lam, 2023). For reference, Willow Garage was a robotics research lab known for developing open-source projects and, at the time, standard robotics software (Vance, 2014). The repository contains sound samples from 13 categories, ranging from human, vocoded linguistic speech to synthetic, non-linguistic audio such as beeps and whistles (Lam, 2023). A total of 309 sounds were available, out of which 38 were filtered out due to linguistic content. Later, from each of the 13 categories, a maximum of 8 sounds were chosen at random (if less than 8 sounds remained in a certain category, all of the samples from that category were chosen). That resulted in 84 sound samples from this library being included in this project.

Bremen Emotional Sound Toolkit

The Bremen Emotional Sound Toolkit (BEST) is a set of nonverbal auditory emblems designed for robots as part of the EMOTE Project (Kappas et al., 2014). The corpus consists of sounds recorded by 9 experts. Each expert was asked to synthesize 60 sounds using a tablet equipped with a synthesizer. Out of the 60 sounds, 20 were based on acted speech, and 40 were based on 'emotion' sounds. The emotions included anger, disgust, enjoyment, fear, interest, sadness, shame, and surprise, and each of them was recorded in both 'low' and 'high' intensities. The recordings were later evaluated in a set of online tests (Kappas et al., 2014).

From the 408 files available in the BEST corpus, 238 were labelled (labels included the emotion and its intensity). For this project, 10 sounds per emotion were selected, with an equal distribution between high and low intensity levels, resulting in a total of 80 sound samples from the BEST corpus included in the listening test.

Gibberish Speech

Finally, it was decided to also include samples from gibberish speech i.e., speech sounds that do not carry any meaning or sense (Yilmazyildiz et al., 2011). Yilmazyildiz et al. (2011), created the Emotional Gibberish Speech Database (EMOGIB) that contains recordings of a single actress portraying both, a neutral state, and six emotions (anger, disgust, fear, happiness, sadness, and surprise). From each emotion, 11 sounds were chosen at random.

Additionally, to incorporate sound samples of Gibberish speech that were assumed to sound more robotic, recordings from the ROBOGIB corpus were included. The corpus was created by Ermers (2023) who adapted sounds from EMOGIB by Yilmazyildiz et al., 2011.

The pitch of the samples was shifted by two semitones and a delay of 50ms was added, based on the work of Wilson and Moore (2017). Similarly to EMOGIB, 11 sounds per emotion were chosen at random.

Semantic-free utterances used in human-robot interaction have been classified by Yilmazyildiz et al. (2015). According to this classification, two main types: gibberish speech (GS) and non-linguistic-utterances (NLUs) can be distinguished. Moreover, Yilmazyildiz et al. (2015) also distinguish paralinguistic utterances (PU) and musical utterances (MU). GS and PU are both vocalizations that resemble human speech. Nevertheless, GS is characterized by meaningless speech sounds, whereas PUs are non-speech vocalizations such as laughs or sighs. On the other hand, NLUs and MUs both do not resemble natural speech. NLUs comprise of synthesized *beeps, squeaks, and whirrs* (Yilmazyildiz et al., 2015, p.65). MUs are primarily driven by musical theory, distinguishing them from NLUs.

Based on those characteristics, the sounds included in this research can be classified. Sound samples coming from the Montreal Affective Voices corpus are PUs. Therefore, the sound samples from the Montreal Affective Voices corpus that have been adapted to sound more robotic (MONTREAL ROB.) can also be classified as such. Likewise, certain sounds of vocoded actors from the Willow Garage HRI Sound Library are also PUs. On the other hand, EMOGIB and ROBOGIB both contain only samples of GS. Furthermore, the synthesized emotional sounds in the BEST corpus can be classified as NLUs together with certain synthesized sounds included in the Willow Garage HRI Sound Library. The summary of all the corpora used together and the types of utterances they contain is presented in Table 1.

Table 1: Final number of sound samples included in the listening test per corpus

Corpus	**#S**	**#DS**	**Type of Utterance**	**Source**
MONTREAL	82	10	PU	(Belin et al., 2008)
MONTREAL ROB.	76	10	PU	(Belin et al., 2008)
WILLOW	84	13(1)	PU & NLU	(Lam, 2023)
BEST	80	9(2)	NLU	(Kappas et al., 2014)
EMOGIB	77	1	GS	(Yilmazyildiz et al., 2011)
ROBOGIB	77	1	GS	(Ermers, 2023)
total	**476**			

#S – number of sounds included #DS – number of distinct speakers in the corpus

(1) - 13 folders of sounds each one being labeled as from a different source (2) - 9 different experts who each have used the same synthesizer to create sounds

Other Considered Sources

Other considered sources included the VENEC corpus of vocal emotion expressions and the variably intense vocalizations of affect and emotion corpus (VIVAE) (Holz et al., 2022; Laukka et al., 2010). The prior one was excluded due to identified background noise in the recordings, especially when the robotic filters were applied to them. The latter one, however, was not clearly labelled, and hence it was decided to also exclude it.

Moreover, it was considered to employ robotic voice samples from media. Nevertheless, after thorough research, it was discovered that there is a very limited set of such samples publicly available. All research papers that were identified to use voices from movies did not have the corpora published or available (Kriz et al., 2010). From websites with publicly available robotic samples from movies, VoicyNetwork and 101Soundboards were considered but ultimately not included due to the sound samples' limited number, variation, and qualities (101soundboards, n.d.; Network, n.d.).

4.1.3 Measures

Both, the *roboticness* and *suitability* to social robots of each sound sample were measured on a 7-point scale. For the scores of *roboticness*, the scale ranged from *Extremely human-like* to *Extremely robot-like*. Similarly, for suitability to social robots, the scale ranged from *Extremely unsuitable* to *Extremely suitable*. Subsequently, perceptions of the causes of *roboticness* and the associations that participants had with robots were measured in the form of open-ended response questions. The questions were formulated in the following way: (i) *What influenced your ratings about the roboticness of the sounds?* for gathering perceptions of the causes of *roboticness*, and (ii) *What robot(s) did you have in mind when responding to this survey?* for assessing participants' associations with robots.

4.1.4 Setup and Procedure

Data was collected in the form of an online listening test. For this purpose, Qualtrics (2005), an online survey software, was employed. Following an opening statement, participants were shown a disclaimer advising them to complete the survey in a quiet environment or use headphones to prevent inaccuracies caused by intervening noises. Henceforth, participants were asked for their consent to participate in the study and demographic-related questions were posed. This series of questions asked about gender, age, highest level of education achieved or currently pursued and cultural identity. Additionally, respondents were asked to self-report their level of familiarity with robots and provide an explanation for the source of that familiarity through an open-ended question.

The following section of the survey included the listening test. Each participant was presented with 50 sounds that were chosen at random (employing the *evenly randomized sample* feature of Qualtrics) from the sample of 476 sounds in the corpus. Short instructions and a note asking participants to pay attention to the voice quality of the sound sample and not to the expressed emotion were also included (see Figure 3a). This disclaimer was incorporated as some of the corpora used were specifically designed to contain emotional speech.

Per each of the sound samples, two questions were posed: one regarding the *roboticness* (see Figure 3b) of the voice of the recording and the other concerning the perceived suitability of that voice to a social robot (see Figure 3c). Additionally, a definition of a social robot was provided in a slightly smaller font. This definition was present for every question of this type (see Figure 3a). Each sound sample was presented on a separate page of the survey and the participants could not have omitted any of the listening test questions.

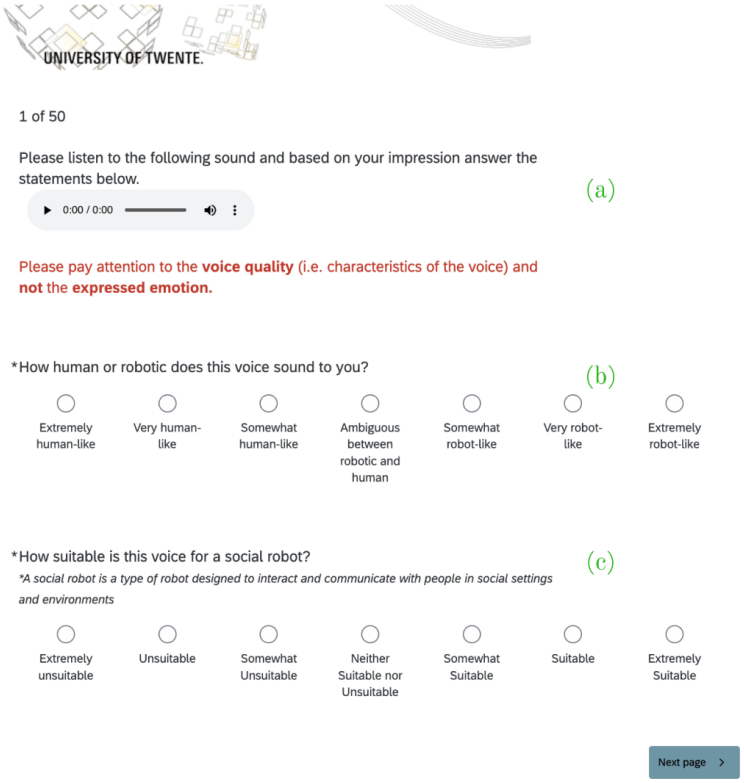


Figure 3: Qualtrics survey: listening test page

After answering questions about 50 sounds, participants were presented with two open-ended questions, the first one concerning factors that influenced their ratings about the *roboticness* of the sounds and the second one about robots that they had in mind when answering the questions (see Figure 4).

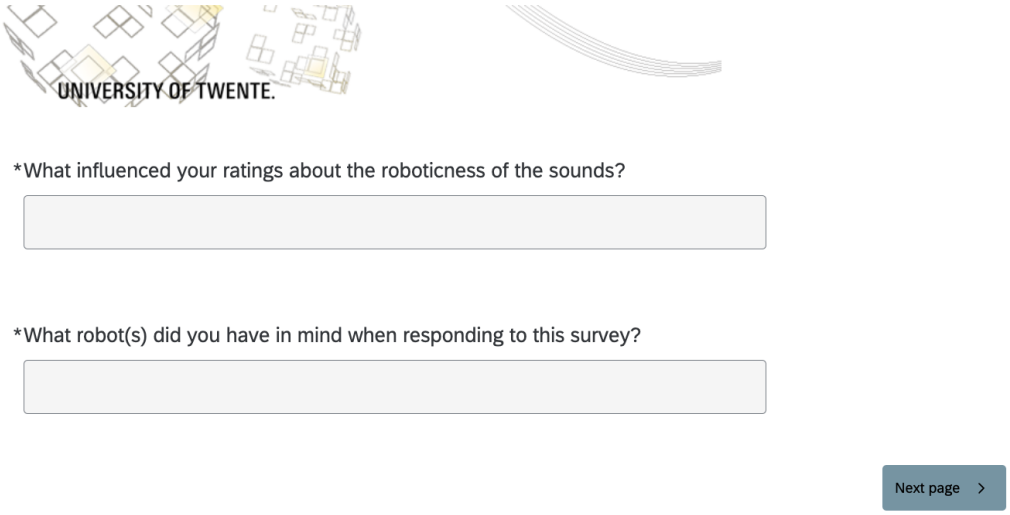


Figure 4: Qualtrics survey: open-ended questions

Finally, a closing statement was presented and the response was recorded. All incomplete responses were set to be discarded after an hour using the *Incomplete survey responses* feature in Qualtrics (Qualtrics, 2005).

4.1.5 Data Analysis

First, in order to answer RQ 1.1, the scores of *roboticness* and suitability for social robots for each sound sample were averaged. Then, a simple linear regression analysis was performed to see if there was a linear relationship between those two types of scores. To do that, `scikit-learn` and `statsmodels` libraries were employed (Pedregosa et al., 2011; Seabold & Perktold, 2010).

As mentioned above, participants were also asked to respond to two open-ended questions – one regarding the factors that influenced their ratings of *roboticness*, and the second asking for robots that participants had in mind when responding to previous questions. The former question explored participants' perceived factors influencing their ratings of *roboticness*, thus addressing RQ 1.2. The responses to that question were analyzed using thematic analysis, following the method proposed by Braun and Clarke (2012). After familiarization with the responses, the initial codes were generated and textual data coded. This was followed by grouping similar codes together to identify themes in the participants' answers. ATLAS.ti version 24.1.0 for Mac (2023), a qualitative data analysis software, was employed to ease and structure the coding process. The responses to the latter one were analysed quantitatively to extract the most common robot that people associate the word 'robot' with and hence answer RQ 1.3.

4.2 Results

4.2.1 Summary of the Ratings of *Roboticness* and Suitability

As can be inferred from the two histograms that represent the frequency of each rating of *roboticness* (Figure 5) and suitability (Figure 6) shown below, the dataset of scores was not entirely balanced. For *roboticness* (Figure 5), where 1 is the *Extremely human-like* rating and 7 is the *Extremely robot-like* rating, it is clearly visible that participants often rated sounds as with the highest *roboticness* ratings. For the ratings of the suitability of voices to social robots (Figure 6), 1 is the *Extremely unsuitable* and 7 is the *Extremely suitable* rating. The distribution of those ratings appears to be more balanced compared to the *roboticness* ratings. Nevertheless, scores of extreme applicability to social robots were clearly rarely given.

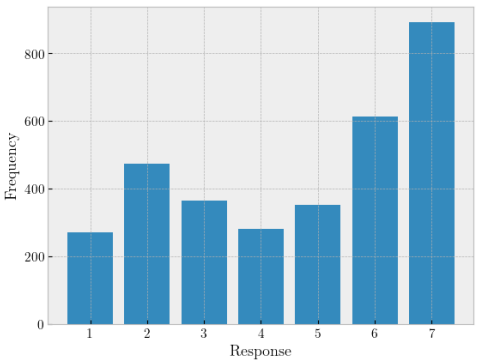


Figure 5: Histogram of ratings of *roboticness* scores

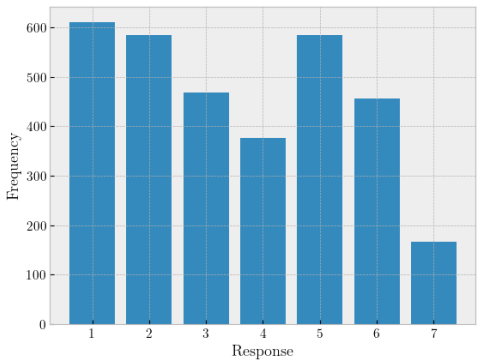


Figure 6: Histogram of ratings of suitability scores

Furthermore, to analyze response variability, standard deviations of responses per sound per each of the ratings were computed. The average standard deviation for *roboticness* ratings was around 1.0, while for suitability ratings it was around 1.6. This suggests that the responses were more variable for suitability compared to *roboticness*.

As explained in the Section 4.1.2, 6 different *types* of corpora were included in the survey. Some of them consisted of purely human sounds (namely, EMOGIB and the Montreal Affective Voices) and others from either synthetic voices (EMOTE and partly Willow HRI Sound Library) or adapted human sounds (ROBOGIB, Montreal Affective Voices Robotic and partly Willow HRI Sound Library). Therefore, evaluating whether certain corpora were evaluated significantly differently than others is crucial to gaining valuable insights.

Corpus	**Mean Avg. Score Roboticness**	**Std. Avg. Score Roboticness**	**Mean Avg. Score Suitability**	**Std. Avg. Score Suitability**
MONTREAL	2.336053	0.542191	4.622367	0.794005
EMOGIB	2.431329	0.628509	4.811673	0.711301
ROBOGIB	4.974959	0.569196	3.235240	0.575234
MONTREAL_ROB.	5.587981	0.656206	2.655882	0.692555
WILLOW	6.108579	0.919970	2.982285	0.882778
BEST	6.459524	0.414882	2.941240	0.777189

Table 2: Comparison of mean and standard deviation scores of avg. suitability and *roboticness* ratings per corpus

Table 2 demonstrates that the BEST and Willow HRI Sound Library Corpus were rated as the most robotic. Those two corpora largely contain synthetic sounds. On the other hand, the two purely human sets of sound (EMOGIB and the Montreal Affective Voices) are rated as the most human, which indicates that people generally can distinguish between synthetic and human voices. Having ROBOGIB and Montreal Affective Voices Robotic rated slightly lower than the two corpora containing only synthetic sounds might indicate

that human voices that have been changed with some voice effects are perceived as quite robotic, yet less robotic than purely synthetic utterances.

Table 2 further shows that the sounds from the two corpora rated the least robotic were also rated as the most suitable for a social robot, with the mean average scores in the range of 4.6 to 4.8 (out of 7.0). On the other hand, the sounds that were rated less suitable all had similar mean average scores in the range of 2.6 to 3.0 (out of 7.0). Moreover, ROBOGIB was rated higher than all of the other robotic voices and EMOGIB was rated the highest overall, which might indicate that participants perceived speech with some semantic content as more applicable to social robots than simply non-linguistic utterances.

4.2.2 Relationship between 'Roboticness' and 'Suitability'

The figure below presents a scatterplot illustrating the relationship between the average rating of *roboticness* and suitability, along with the fitted line of regression. Just by looking at the graph, the clear linear relationship between the scores of *roboticness* and suitability is visible.

The linear regression analysis revealed a statistically significant ($p < 0.05$) relationship between the *roboticness* of a voice and its suitability to a social robot with $\beta = -0.495$, indicating a negative relationship, i.e. the more robotic given voice is the less applicable it becomes. The R-score (R^2), which explains the proportion of variance in the suitability that can be explained by the *roboticness* is $R^2 = 0.601$ suggesting that the linear regression model has a satisfactory fit and can explain a significant portion of the relationship between variables.

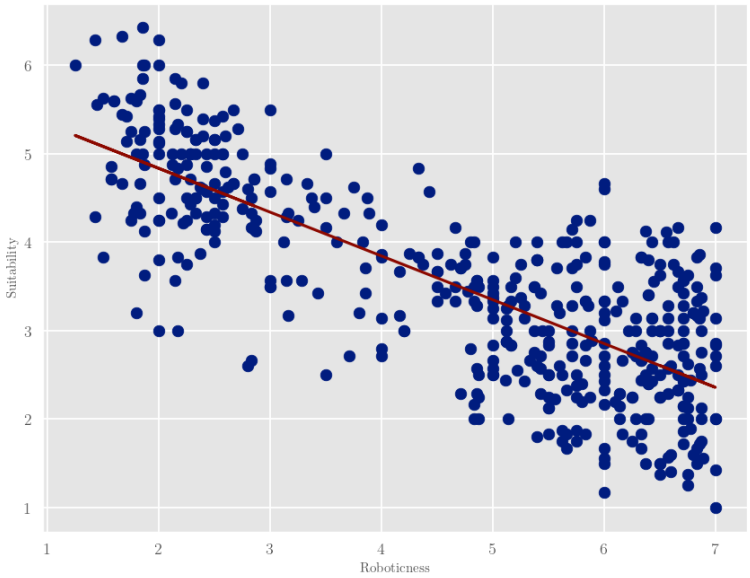


Figure 7: Linear relationship between the *roboticness* of a voice and its suitability for a social robot

4.2.3 Factors that Influence Perceived *Roboticness*

Qualitative, thematic analysis was conducted to analyse the factors that participants of the study self-reported as influencing their ratings of *roboticness*. Eight main themes were identified in the analysis. 1. *Vocal Variation*, 2. *Layered Voices*, 3. *Pleasantness of Vocal Delivery*, 4. *Lignugistic Properties of the Voices*, 5. *Specific Vocal Features*, 6. *Machine-like Vocality* 7. *Sound Quality* 8. *Imageability*

Vocal Variation theme encompasses parts of responses that referred to the intonation of the sounds. Participants often mentioned *tone* of the voice as well as *modulation*, *intonation* and *monotonicity*. Participants mentioned rating sounds that had more *dynamic changes in the sound's tone* and in which *frequency was modulated naturally* as less robotic and voices that were *monotonous* as robotic.

Layered Voices theme was identified from responses that mentioned *multiple speakers*, sounds in the *background* or *echo*. It is likely that echos or sounds that were modified by adding an overlaying layer of delayed sound affected the perceived *roboticness* of the voice.

Pleasantness of Vocal Delivery theme refers to responses that mentioned rating sounds that were *unsettling*, *offputting*, *irritating*, *annoying* as robotic and *pleasant* or *comfortable* as human-like. Such qualities of the voices were mentioned in total ten times indicating that it was an important factor that people took into account when rating the sounds. Nevertheless, it is important to note that as some of the sounds contained emotional speech, the unpleasantness of the sounds might have come from the negative emotion rather than the timbre of the voice itself, which was also indicated by a few of the participants.

Lignugistic Properties of the Voices theme refers solely to responses that mentioned rating speech-like sounds as less robotic. Three respondents mentioned taking into account the similarity of the sound to a word rather than just a *beep*. This theme might hence refer to the inclusion of gibberish speech samples and shows that including linguistic properties of speech lowers the perception of the *roboticness* of a sound.

Specific Vocal Features theme refers to concrete prosodic features mentioned by participants, i.e. *frequency*, *pitch*, *volume* and *amplitude*. The last one referred to voices with *very high-amplitude* being rated as robotic.

Machine-like Vocality theme was identified purely of responses describing robotic speech as *metallic* or *mechanical*.

Sound Quality theme refers primarily to the sounds being *clear* or *clean*, properties mentioned in total six times by the participants. Moreover, one participant mentioned *cracking* of the voice as a factor.

Imageability theme was identified from the descriptions of the sounds related to how *familiar* or *natural* given voice sounded to participants.

4.2.4 Associations with Robots

When asked about robots that participants had in mind when rating voices on the *robotic-ness* and suitability to social robots scales, two main categories of answers can be found. People either referred to existing, concrete examples of artificial agents or to specific applications or functionalities of robots.

When referring to concrete examples of robots, participants often mentioned robots from media. The most common examples were robots from *Star Wars* with R2-D2 being referred to 6 times, C-3PO 2 times, BB-8 1 time and *Star Wars* alone 4 times. The second robot from media, mentioned 4 times in the responses, was WALL-E from the movie with the same name. Moreover, participants often referred to voice assistants, in particular to *Siri* and *Alexa* who each was mentioned 3 times. Respondents also brought up some singular examples of real-life robots such as *Photon*, *Nao*, *Furhat* or *Kerfuś*. Table 3 displays a list of concrete examples of artificial agents mentioned by participants.

	Artificial Agent	No.: mentioned
Movies	R2-D2 (*Star Wars*)	6
	WALL-E *(WALL-E, 2008)*	4
	C-3PO (*Star Wars*)	2
	Eve *(WALL-E, 2008)*	1
	BB-8 (*Star Wars*)	1
	TARS *(Interstellar)*	1
	Robby the Robot *(Forbidden Planet)*	1
	Marvin *(The Hitchhiker's Guide To The Galaxy)*	1
	Funnybot *(South Park)*	1
Real-life	Alexa	3
	Siri	3
	Furhat	1
	Kerfuś	1
	Nao	1
	Photon	1

Table 3: Artificial agents examples mentioned by participants

On the other hand, many respondents mentioned certain applications of the robots they considered. Among the most popular applications were waiter robots, highlighted 4 times, care robots, particularly in nursing homes and hospitals, mentioned 3 times, along with chatbots and toys, each mentioned 3 times.

5 Study II: Speaker Embeddings and Vocal Parameters Analysis

5.1 Methodology

As mentioned before, this study employed two methods of feature extraction – speaker embeddings and eGeMAPS features. Each of those extracted features was then analyzed in relation to both *roboticess* and suitability to social robots ratings. The primary algorithm employed to analyse whether relationship between either of those features was linear regression analysis with k-fold cross validation. Moreover, dimensionality reduction and feature selection algorithms were employed to improve the analysis.

5.1.1 Speaker Embeddings

Speaker embedding representations were extracted using the `WavLM Large Model` trained model that has been pre-trained on 70,000 hours of English linguistic speech and 24,000 hours of multilingual speech (Chen et al., 2022; "microsoft/wavlm-large · Hugging Face", n.d.). The model weights were loaded using the `from_pretrained` method with default parameters. The model consists of a convolutional neural network (CNN) as a feature encoder that extracts features from the raw audio and a transformer encoder network that captures contextual information and dependencies among these features, creating a representation of the audio file (Chen et al., 2022). The `WavLM` contains 24 transformer encoder layers and, in total, 316.62M parameters. Moreover, the sampling rate at which the audio files should be digitized was 16,000 Hz. This required the sound samples to be resampled to match this rate before being put into the model. In order to extract speaker embeddings in the form of 512-dimensional vectors `WavLMForXVector` feature extractor head was employed (Chen et al., 2022; "microsoft/wavlm-large · Hugging Face", n.d.)

Subsequently, to gain initial insights into the data and explore potential patterns related to RQ 2.1 and RQ 2.2, the extracted feature vectors were visualized before proceeding with further analysis. Visualizing speaker embeddings can be useful to understand the patterns and anomalies happening in the data. However, since speaker embeddings are high-dimensional vectors, visualizing them without any modifications is not possible. That is why research papers investigating speaker embeddings often utilize dimensionality reduction techniques with t-SNE being a commonly utilized method (de Seyssel

et al., 2022; Stan, 2022; Ulgen et al., 2024). T-SNE is a non-linear dimensionality reduction technique that maps high-dimensional data to a lower-dimensional, in this case 2-dimensional, space, aiming to preserve the local structure of the data points (van der Maaten & Hinton, 2008). This makes it suitable for visualizing the extracted speaker embeddings, as mentioned before, are 512-dimensional vectors. In order to perform t-SNE, `TSNE` class in the *Manifold Learning* module available through the `scikit-learn` library was employed (Grisel et al., 2024; Pedregosa et al., 2011). The perplexity was set to 7 due to a relatively small dataset and otherwise run with the default parameters, i.e. number of iterations for the optimization set to 1000 and Principle Component Analysis (PCA) initialization.

5.1.2 Acoustic Parameters

As mentioned before, to extract vocal features from the sound samples in a systematic manner, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was employed (Eyben et al., 2016). The set is available through the `OpenSMILE` toolkit and consists of 25 'Low-Level Descriptors' (LLD). Out of them 18 are present in the minimalistic set and 7 were added to the extended set (Eyben et al., 2016; Eyben & Schuller, 2010). The extended, precisely, the `eGeMAPSv02`, version of the set was chosen as 4 of those added parameters belong to the mel-frequency cepstral coefficients (MFCC), which are often used to analyse the timbre of a voice (Eyben et al., 2016). Similarly to the speaker embeddings extraction, eGeMAPS features were extracted separately for each sound sample and then compared to the average of either *roboticness* or suitability to social robot rating of that sound sample.

5.2 Analysis

5.2.1 Speaker Embeddings

In order to inspect whether speaker embeddings encode the *roboticness* and suitability of a voice to a social robot, a machine learning model was applied similarly, as has been done by Stan (2022). Since the scores of both factors are continuous rather than discrete, regression models were the most applicable to use.

As the dataset is relatively small, containing only 476 data points, and the speaker embeddings are high-dimensional (512-dimensional) vectors, a dimensionality reduction technique was applied to prevent the model from overfitting. Namely, the Principal Component Analysis (PCA) was employed to reduce the size of the embedding vectors to 50 dimensions.

To further mitigate the limitation coming from a relatively small dataset (corpus), 5-fold cross-validation with a random state set to 42 was employed. For each fold, a linear regression model was fit, and performance metrics – R^2, mean squared error and mean absolute error, calculated. To conduct this analysis, functions from the `scikit-learn` library were employed (Pedregosa et al., 2011).

5.2.2 Acoustic Parameters

In order to analyse which acoustic feature contributes to the voice sounding the most robotic and most suitable for a social robot, a simple linear regression analysis was conducted separately for each of the ratings. Furthermore, similarly to Bosland (2022), the Recursive Feature Elimination (RFE) was used to assess which features are most important in making a voice sound more robotic or suitable for social robots. The RFE algorithm removes less important features in an iterative process to create a subset that maximizes the predictive accuracy of a given prediction model (Avcontentteam, 2023). In a way, it is similar to the PCA that was used for speaker embeddings dimensionality reduction. Nevertheless, RFE was chosen due to its ability to preserve the interpretable results, i.e., to know the exact names of the parameters that contribute to a given rating.

RFE with a linear regression model as an estimator was conducted iteratively for the number of features between 1 and 88 (88 being the total number of features from eGeMAPS). As in the analysis of the speaker embeddings, to mitigate the limitation coming from a relatively small dataset (corpus), 5-fold cross-validation with a random state set to 42 was employed. Within each iteration of RFE, the model was trained on 4 folds of the data and evaluated on the remaining fold using R^2 and negated mean squared error. Both of the scores were averaged after each iteration, resulting in a more robust estimate of the model's performance. For all, the linear regression model, the RFE and the k-fold evaluation `scikit-learn` library was employed (Pedregosa et al., 2011).

5.3 Results

5.3.1 t-SNE Vizualization

Figure 8 displays the t-SNE visualization plot of speaker embeddings coloured by the scores of *roboticness*. As t-SNE clusters similar data points together, the ratings for each sound sample were averaged to the nearest integer number and hence clustered into groups of possible ratings, where 1 represents the *Extremely human-like* and 7 *Extremely robot-like* (van der Maaten & Hinton, 2008). Note that there were no sound samples with a rounded average rating of 1, which is why they are not visible in the figure.

It is clearly visible that, in most cases, embeddings of sounds with similar ratings are clustered together, i.e. 'darker' dots tend to be placed in clusters with other 'darker' dots and likewise for the 'lighter' ones. This makes it probable that the speaker embeddings can, to a certain extent, be predictors of the *roboticness* of the sound sample.

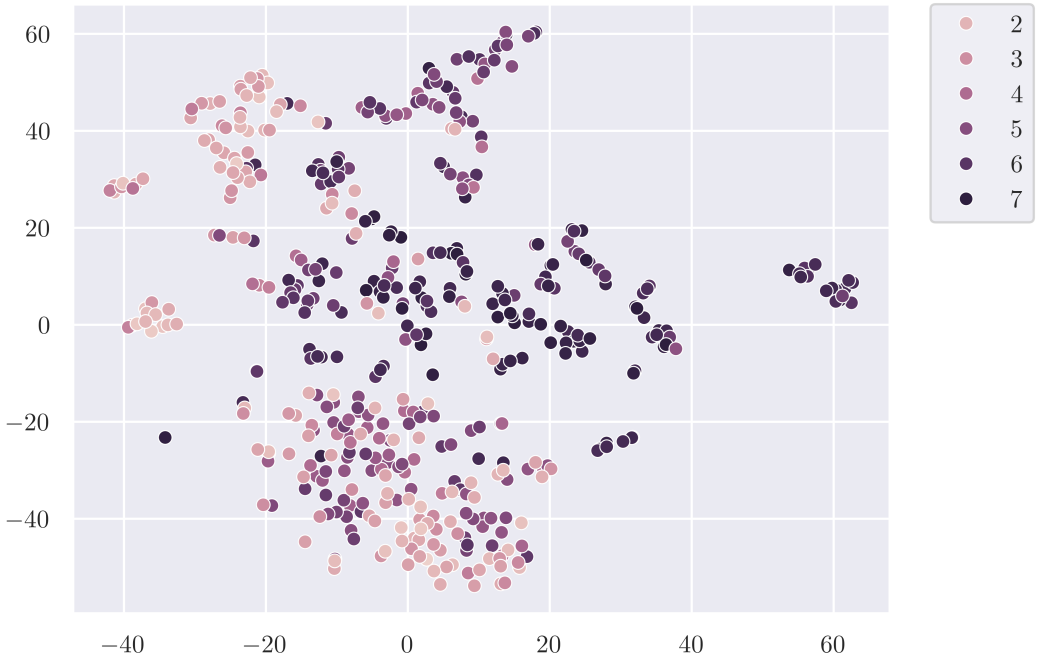


Figure 8: t-SNE Visualization Plot of Speaker Embeddings Colored by Rounded Avg. *Roboticness* Score

Figure 9 depicts the same t-SNE plot of speaker embeddings as in Figure 8 but coloured by the ratings of the rounded average ratings of the suitability of a given sound for a social robot. Similarly to Figure 8, the clusters visible on the plot typically showcase a prevalence of either darker or lighter-coloured data points. However, this characteristic seems to be less evident for the 'suitability' plot as compared to the *roboticness* one.

Furthermore, the reversed relationship between the ratings of *roboticness* and suitability to social robots is explained in the previous section is also noticeable. For example, data points on the upper right side of the graphs are primarily coloured in dark purple on Figure 8 and in light purple on Figure 9.

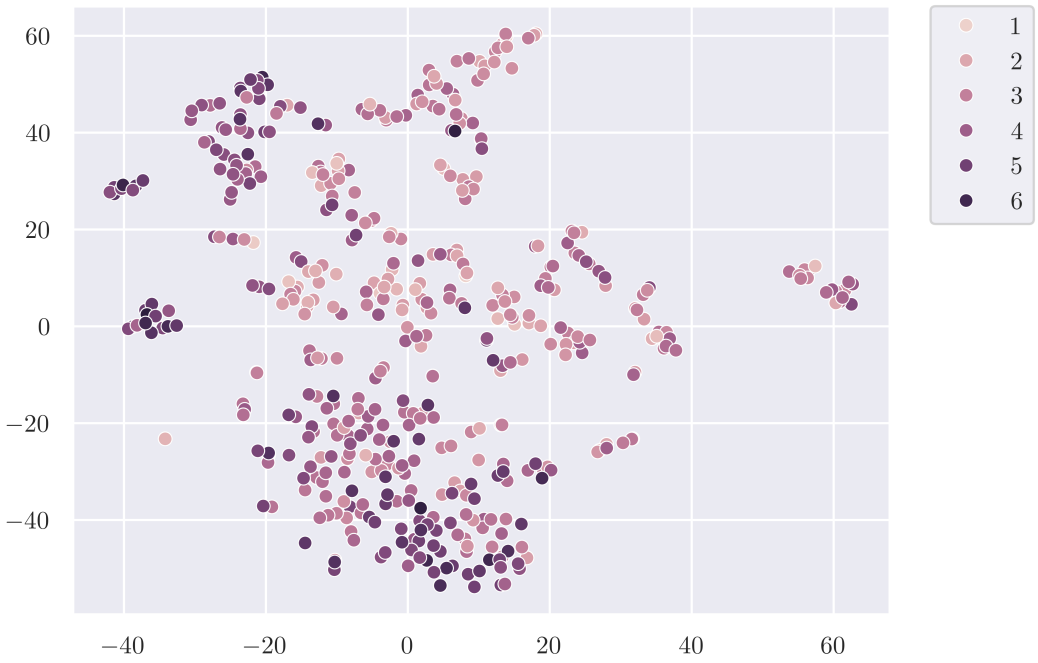


Figure 9: t-SNE Visualization Plot of Speaker Embeddings Colored by Rounded Avg. 'Suitability' Score

Please note that t-SNE focuses on preserving local structures and not global ones. This means that larger distances between clusters do not necessarily mean larger distances between the data points in those clusters (van der Maaten & Hinton, 2008).

5.3.2 Predictions Based on Speaker Embeddings

The table below displays the performance metrics of linear regression with k-fold cross-validation for predicting the scores of *roboticness* and suitability of voices to social robots. Firstly, the model performed significantly better at predicting *roboticness* scores compared to the suitability scores. This is evident from the much higher R^2 value for *roboticness* predictions. However, it is surprising to see that the average error metrics, i.e. mean squared error and mean absolute error, are lower for the suitability scores.

	Roboticness	Suitability
R^2	0.552250	0.291118
MSE	1.423387	0.897338
MAE	0.968920	0.742872

Table 4: Evaluation of Linear Regression with k-fold Cross Validation on Speaker Embeddings

5.3.3 eGeMAPS Features: Simple Linear Regression

Simple linear regression did not yield definite results. The most relevant feature was *slopeUV0-500_sma3nz_amean* suggesting weak positive associations for both ratings. For the ratings of *roboticness* the linear regression resulted in $R^2 \approx 0.30$, mean squared error of ~ 2.25 and a statistically significant slope of $\beta \approx 17.14$. For suitability to social robots ratings, however, the R^2 score was slightly lower with $R^2 \approx 0.24$, mean squared error ~ 0.97 and a negative slope $\beta \approx -9.58$. Tables with all eGeMAPS features sorted by R^2 for both ratings are included in Appendix A.

5.3.4 eGeMAPS Features: Recursive Feature Elimination

The results of the RFE in combination with the k-fold evaluation revealed that the combination of a smaller number of features generally causes the model to underfit. For the prediction of the scores of *roboticness* the value of R^2 continues to rise until 9 features are selected. Later, the R^2 score settles in the range of 0.4 to 0.5. That remains to be the case until the number of features is 52 and R^2 scores become negative indicating overfitting. Somewhat similar results are present for predicting the suitability of voices for social robots with the difference being that the best R^2 scores are significantly lower for the predictions of the suitability. Table displaying the R^2 and mean squared error values for both *roboticness* and suitability per each number of features selected is present in the Appendix B.

The features that are primarily selected to predict *roboticness* are related mainly to the fundamental frequency (f0), indicating pitch as an important factor, loudness, spectral flux, MFCC, jitter (deviations in individual consecutive f0 period lengths) and shimmer difference of the peak amplitudes of consecutive f0 periods). The list of the top 35 features predicting *roboticness* is present in the Appendix B. Similar features tend to also predict the suitability of a voice to a social robot further which is likely associated with the fact that *roboticness* is strictly correlated with the suitability of a voice to a social robot as explained in the previous section. This number of features was included as it resulted in the highest R^2 of approximately 0.49. The list of the top 39 features predicting the suitability of a voice to a social robot is also present in the Appendix B. The top 39 were chosen with the same motivation as the *roboticness* score as it resulted in the highest R^2 of approximately 0.29.

5.4 Discussions and Limitations

Despite certain limitations that cannot go unmentioned, the results of both studies provide insights into how people perceive robotic voices and how those perceptions could be modelled. Moreover, the outcomes of Study I support certain assumptions made in Study II and complement its results.

5.4.1 Relationship Between *Roboticness* and Suitability to Social Robots

Possibly the most definite result of this study is related to the RQ 1.1. As hypothesized, the relationship between the *roboticness* and suitability to social robots has proven to be negative when conducting the simple regression analysis. $\beta = -0.495$ indicates that, on average, as the scores of the suitability to social robots increase by one point, perceived *roboticness* tends to decrease by 0.5 points. Moreover, the $R^2 = 0.601$ implies that *roboticness* can explain 60% of the variance in the suitability to social robot scores. Moreover, this relationship could have also been observed in the t-SNE visualization graphs in Study II. Data points represented as dark dots on the t-SNE visualization plot of speaker embeddings, coloured by rounded average *roboticness* scores (Figure 8), are depicted as light dots on the plot where data points were coloured by suitability to social robots scores (Figure 9). This highlights the opposition between these two sets of scores.

Those findings indicate that people generally think that social robots should sound more human-like, supporting the results of the studies by Kühne et al. (2020), Gurung et al. (2023) and Schreibelmayr and Mara (2022). Nevertheless, the fact that participants were not provided with any visual or physical representation of a robot must be acknowledged and can cause certain limitations. For example, research by McGinn and Torre (2019) has shown that people form a certain mental image of a robot when they hear a voice. Participants in that study were asked to match visuals depicting robots with voices that differed, among others, in naturalness. This variable was then shown to significantly affect the participants' choices of the most suitable robot for that voice. Therefore, given the diverse range of robots pictured by participants in this survey, as explained in the Section 4.2.4, it is plausible that perceptions of the appropriateness of voices for social robots may have varied and would have been different if participants were provided with images or concrete examples of social robots.

5.4.2 People's Perception of *Roboticness*

Guided by RQ 1.2 – *What perceived factors influence individuals' assessments of roboticness of voices?* – the investigation into what this *roboticness* means to people was done by asking participants of the survey about the factors that influenced their ratings of the sounds. Results of the thematic analysis of those responses are, to a certain extent, similar to the ones of a study by Kühne et al. (2020). The *Imageability* theme is present in both studies, suggesting that people perceive voices as robotic when they are not able to imagine them as those of a speaking human. Some overlap can also be found between the *Sound* theme identified by Kühne et al. (2020) and the *Machine-like Vocality* theme identified in this study. In the latter one, participants often described voices as *metallic* or *mechanical*, while in the prior one descriptors such as *choppy* and *technical* and *metallic* were present. Notably, the term *metallic* appeared in both studies.

Furthermore, the *Vocal Variation* theme from the present study also intersects with the *Intonation* theme from the study by Kühne et al. (2020) indicating that a rhythm or modulation of a given sound is an important factor contributing to *roboticness*. This theme can also be interpreted in relation to the vocal features identified in the Section 2.2. More specifically, the theme could be pointing at the higher frequency spectrum of robotic voices as compared to the human ones.

Even more deliberate connotation with the methods of creating robotic voices from Section 2.2 can be found in the *Layered Voices* theme. That theme centres around the concept of an *echo*. This effect is often achieved through delays applied when modifying a human voice to sound more robotic using filters such as Voxal (NCH Software, n.d.). The effects of robotic filters might also be explaining the *Vocal Features* theme identified, which consisted of concrete vocal parameters that participants were able to identify. The features most commonly referred to by participants were frequency and volume-related ones, which are in line with the amplification and pitch shift modifications applied to some of the sounds in the corpus (see Section 4.1.1). Furthermore, the *Sound Quality* theme hints at additional properties of pitch that can cause a certain voice to sound more robotic, for example, pitch shimmer that could contribute to impressions of *unclear* or *cracking* voices.

Additionally, the thematic analysis revealed that the *Linguistic Properties of the Voices* of voices play a crucial role in them being perceived as less robotic. It might be inferred that robots who speak human language should be perceived as less robotic. This finding is further substantiated by the average ratings of sounds per corpus presented in the Section 4.2.1. Notably, the ROBOGIB corpus received the lowest average ratings for *roboticness* compared to all other corpora containing synthetic or sound samples adapted to sound more robotic.

Finally, the thematic analysis revealed a theme related to the pleasantness of a given voice. As much as it is a quite subjective impression, the results suggest that there might exist a relationship between voices sounding pleasant and them sounding more natural. This finding then also further supports the hypothesized relationship between *roboticness* of a voice and its suitability for a social robot as it is logical to assume that social robots should have voices that are pleasant to humans.

5.4.3 Associations with Robots

To evaluate whether robots from media can offer insights into understanding the perception of *roboticness* of voices, RQ 1.3 – *What are people's associations with robots?* – was posed. To address it, a question about the robots that respondents had in mind when answering the survey was added. Based on the responses, it is evident that people often still associate robots with fictional characters from the media. This is in accordance with the results of a study by Kriz et al. (2010) indicating that the effect of media on

the perception of robots is still significant despite the advancements in technology over the last decade. On the other hand, a significant number of respondents mentioned existing, real-life examples of artificial agents, with the most common ones being virtual voice assistants. This indicates that such technologies are becoming embedded and recognized in today's society. Interestingly enough, a few examples of research robots were mentioned.

It is crucial to consider that the results may have been influenced by the following two factors: (i) the sounds presented to participants were non-linguistic linguistic, which might have affected the resemblance of the sounds with certain to robots from media, for example, *R2-D2* from *Star Wars*; (ii) above 50% of the respondents reported having at least some level of familiarity with robots. Many respondents mentioned such familiarity coming from work or university, which is expected as the survey has been distributed among people associated with the researcher and hence often related to technical studies or university. It is therefore recommended to, in the future, ensure a wider population sample to better assess such associations for society in general.

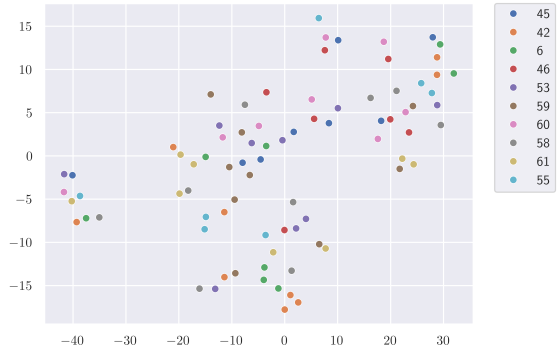
5.4.4 Encoding of the Perception of Robots in Speaker Embeddings

As the thematic analysis revealed, some descriptors of voices related to *roboticness*, such as *Pleasantness* or *Imageability*, may not be easily measurable. It can be deduced that, as hypothesized, it is the timbre or speaker's identity that encompasses this *roboticness*. As such, the analysis of speaker embeddings, assumed to encode the speaker's identity seems to indeed encode information related to *roboticness* and suitability of voices to social robots.

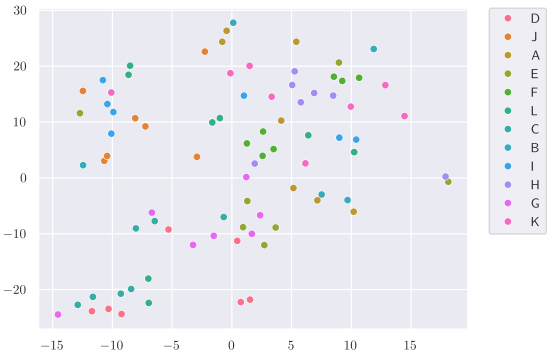
The results of the regression analysis for predicting the roboticness and suitability scores from the speaker embeddings indicate that the model explains approximately 50% of the variance in the *roboticness* scores and 30% of the variance in the suitability scores. Moreover, mean average errors of, respectively, 0.97 and 0.74 suggest that there is a generally acceptable average difference between the predicted scores and actual scores gathered in a survey. Surprisingly, even though for suitability to social robots, the mean squared error score is also relatively low (0.9), for *roboticness* is score is 1.4 suggesting that there might be some outliers in the predictions. This analysis shows that the speaker embeddings do encode information relevant to both ratings to a certain extent providing an answer to RQ 2.1 and RQ 2.2. The reason for the model performing generally worse (lower R^2) for the suitability of voices to social robots might come from the larger variance (as seen in section 4.2.1) for the ratings in the responses, making it a more subjective measure in general. However, it is evident that the model is not performing optimally. Further refinements and the addressing of current limitations are necessary before it can be deemed suitable for practical application.

Perhaps the biggest limitation of this project is that the speaker embedding extraction algorithm has been trained on corpora with linguistic content, primarily in English (Chen et al., 2022). This might make it a less optimal method for the extraction of speaker embeddings from non-linguistic utterances of synthetic voices. On Figure 10 three t-SNE visualization plots of speaker embeddings can be found - first investigating clusters related to corpora used, second investigating whether speaker embeddings are able to distinguish between individual speakers from the Montreal Affective Voices corpus and third investigating whether speaker embeddings are able to distinguish between individual sound libraries of the Willow HRI Sound Library. Such visualizations are not possible for other corpora as they contain either synthetic speech without labelled speakers or, in the case of the Gibberish Speech databases, sounds produced only by one actor.

(a) t-SNE Visualization Plot of Speaker Embeddings Coloured by Corpus



(b) t-SNE Visualization Plot of Speaker Embeddings Coloured by Individual Speakers from Montreal Affective Voices Library



(c) t-SNE Visualization Plot of Speaker Embeddings Coloured by Sound's Library from Willow HRI Sound Library

Figure 10: t-SNE Visualization Plot of Speaker Embeddings Coloured by Corpora and Individual Speakers

The plots clearly demonstrate that the speaker embeddings were not able to clearly distinguish between the speakers of the corpora. Nevertheless, they often clustered sounds belonging to the same corpus together indicating that speaker embeddings encoded some

commonalities in the speaking style of each corpus. Further investigation of the clusters visualized on the t-SNE plots has revealed that it was often the exact utterance type that was clustered together. For example, sound samples of laughing actors' sounds from the Montreal Affective Voices were grouped together (Belin et al., 2008). Similarly, the sounds of bells from the HRI Willow Garage Libary (D and C folders in the repository) were also clustered (Vance, 2014). Retraining the model or creating an entirely new speaker embedding extraction system would therefore be necessary to draw conclusions about the application of speaker embeddings for a similar use case.

An additional constraint associated with this approach is the relatively limited dataset, comprising only 476 sounds. That might have caused some inaccuracies in the prediction models and led to overfitting. It is therefore recommended that such studies be conducted on a larger scale and with a wide selection of sounds and voices.

5.4.5 Analysis of Perceptions about Robotic Voices Through eGeMAPS Features

As the thematic analysis of people's impressions revealed some immeasurable descriptors of *roboticness*, certain themes suggested that participants were able to recognize and point out specific vocal parameters that resulted in robotic-sounding voices such as distorted pitch. Moreover, the associations people hold with robots suggest that robots in media, and consequently robotic voice filters, might help in understanding the perception of robotic voices. Those findings are supported by the vocal features extracted using the eGeMAPS dataset. The features primarily selected by RFE were related to the fundamental frequency (f0) indicating pitch as the most important vocal feature predicting the *roboticness* and suitability of voices to social robots. Additionally, different variations in pitch, i.e. shimmer and jitter, were also important predictors of the *roboticness* which supports the results of the study by Wilson and Moore (2017) and provides answers to RQ 3.1 and RQ 3.2.

Furthermore, timbre-related features, i.e. MFCC and spectral flux features appear to be relevant predictors of the *roboticness* scores. The importance of such features should be explored in future research, ensuring the same conditions for other parameters, e.g., the same pitch and loudness. An important finding, however, is that rather than just one vocal feature, a combination of multiple ones contributes to a voice sounding robot-like. Surprisingly, when analysing the relationship between single features and the *roboticness* score, the mean spectral slope in the low frequency range proved to be predicting the *roboticness* and suitability scores best. It has to be yet noted that, according to, Utsugi et al. (2019), fundamental frequency tends to influence mean spectral slope significantly, making it difficult to draw definitive conclusions about the relationship between it and the *roboticness* of a voice.

The main implications coming from this project are that the more robotic-sounding a voice, the less suitable for social robots. In order to strengthen or decrease this *roboticness* of a voice, one could tweak the f0 and loudness parameters of the voice. Moreover, they could also experiment with robotic filters available online, as they can provide useful knowledge about what will make a voice robotic. For example, adding a delay (echo) to the voice could also increase the *roboticness* of the sound. Finally, to further explore the extent to which a voice sounds robotic, speaker embeddings could be employed. This technology could also be potentially employed to further investigate and model this *roboticness*. However, this implication warrants additional research and refinement of current state-of-the-art technology.

6 Conclusion

Several conclusions can be drawn from both of the studies that were conducted. Perhaps the most definite conclusion of this research is that the more robotic a certain voice is perceived, the less applicable it becomes to social robots. Moreover, the first study has shown that media and robotic filters are valuable sources of knowledge in terms of understanding what makes a voice perceived as robotic and therefore less suitable for social robots. Parameters such as pitch, loudness and delay are often adapted in post-production or in robotic filters. While Study I has shown that those parameters are directly recognizable by humans, Study II proved that they also emerge as most relevant to robotic speech when extracting vocal parameters and comparing them to human judgments of *roboticness*. This suggests that such parameters could be therefore used to quantify and measure the *roboticness* and suitability of a given voice to a social robot. On the other hand, some less measurable descriptors of *roboticness* of a given voice, such as *machine-like vocality* or *imageability* were identified, highlighting that it is actually the timbre or speaker's identity that allows people to distinguish between human and robotic voice. An attempt to extract the characteristics of robotic voice utilizing speaker embeddings has been made. While not entirely conclusive, the results of this research suggest that speaker embeddings can, to a certain extent, quantify *roboticness* and the suitability of voices for social robots. This implies that such technology could potentially be employed to understand the perception of robotic voices in the future. Nevertheless, the current state-of-the-art speaker embedding extraction model does not fully allow for utilising this technology for such purpose as it is meant to extract features from linguistic content rather than short semantic-free utterances that were included in this project.

7 Future Research

To the researcher's knowledge, this work was the first to investigate the perception of robotic speech by employing speaker embeddings. As mentioned before, the method used in this study contains many limitations. Nevertheless, with speaker embeddings being utilized for enhancing text-to-speech technologies, further exploring their usefulness for encoding robotic speech style is recommended (Shaheen et al., 2023; Su et al., 2023; Yu et al., 2022). Future work should focus on (i) extracting speaker embeddings from robotic voices with linguistic content and (ii) retraining or adapting the machine learning model to fulfil the task of extracting the speaker's identity from semantic-free utterances of robotic speech.

Moreover, it is recommended to assess the outcomes of this study in varied real-world settings where individuals engage in physical interactions with robots. As proposed by Dou et al. (2020), the appropriateness of voice design varies based on the specific application of a robot. Hence, exploring the validity of the present findings in more targeted scenarios is encouraged.

8 Contextual Exploration

This section is a requirement by University College Twente to be a part of the Capstone report and its narrative will be guided by the following question: *How could other academic disciplines or fields feed into or profit from this work and how could this work (potentially) be of use to society on a smaller or larger scale?*

8.1 Interdisciplinarity

It is firstly necessary to establish that this project contributes to the fields of human-robot interaction (HRI) and speech processing, both of which are inherently multidisciplinary (Burke et al., 2004; Delić et al., 2019). Moreover, multidisciplinarity is further expressed in this project since it has examined the perception of robotic voices from several perspectives and using various methodologies. A significant portion of this work falls under the umbrella of computational sciences. Various machine learning and data science methods, for example, linear regression models and dimensionality reduction algorithms, have been applied to analyze the gathered data and extract features from the utilized sound samples. Furthermore, the focus on speaker embeddings extraction and analysis in this research highlights its contribution to the field of computational sciences.

However, this project's interdisciplinary nature extends beyond that field. By investigating human perception of sound in the context of robots, the research intersects with the field of behavioural sciences. This is also demonstrated in the qualitative data that has been collected and analysed in Study I. Furthermore, this research utilized knowledge and

concepts from the domain of the natural sciences. Specifically, the science of acoustics, especially in terms of the eGeMAPS features that reflect the physical phenomena of the sound (Eyben et al., 2016).

8.2 Small Scale Implications

Among the few fields that could potentially benefit from this research, the field of robotics is particularly prominent. On a smaller scale, the results of this research provide insights into design principles for robots. Designers could use the knowledge gained from this project to identify and fine-tune the vocal parameters of robots' voices to make them sound more suitable and adjust the level of their robot-like properties. Furthermore, they could potentially employ speaker embeddings as one of the means of indicating the perception of their robot and adapting the *roboticness* of the robot's voice.

Furthermore, this project yields several theoretical implications. As mentioned before, to the researcher's knowledge, speaker embeddings have not been previously used to analyse the properties of robotic speech. The results and conclusions of this research imply that speaker embeddings might potentially be a valuable source of knowledge for understanding such speech after a more thorough evaluation and retraining of the model. Moreover, the technology of speaker embeddings is still being developed (Jakubec et al., 2024). Therefore, this work, to some extent, could benefit further research into speaker embeddings and their applications.

8.3 Large Scale Implications

Taking a broader perspective, this project could make a contribution to the development of robotics overall, in particular social robots. As with every new technology, there are many benefits and threats regarding its development.

As social robots are used in many different applications and fields, this work could, indirectly, contribute to each of them. Mahdi et al. (2022) identify six main applications of social robots present in literature: *service, entertainment, healthcare, education, research and telepresence* (Mahdi et al., 2022, p.6). In the context of robots for entertainment, this project could help improve the sound design and user experience of people using such robots for leisure. When it comes to research robots, the theoretical implications of this project mentioned earlier, could inspire further research and provide certain background knowledge. Moreover, robots for telepresence, defined as robots that can facilitate remote interaction in particular by connecting via audio and/or video are often applied for various purposes. Those include care and education, intersecting with the three main applications of robots left to consider – healthcare, education, and service (Mahdi et al., 2022). Robots in those areas are also often mentioned by Guenat et al. (2022) who consider opportunities and threats to achieving the United Nations Sustainable Development Goals (SDGs)

associated with the development of robotics and autonomous systems (RAS).

Cifuentes et al. (2020) in the review of applications of social robots for healthcare identify several roles of social robots in such a setting. For example, robots can help in therapy, especially in assisting individuals with physical or cognitive limitations, such as those with physical impairments, autism, or other vulnerabilities (Cifuentes et al., 2020; Saleh et al., 2020). Moreover, robots can also serve as companion to individuals who experience loneliness or mental health problems, in particular in eldrely care where there is a shortage of labour (Guenat et al., 2022; Lee et al., 2018; Saleh et al., 2020). Therefore, development of social robots and their improved design that could be supported by this project, may contribute to achieving the SDG 3 – *Good health and well-being* (Nations, 2023).

Similarly, as in healthcare, social robots for education can also perform a range of different tasks (Mahdi et al., 2022). Youssef et al. (2023) identify three distinct tasks in education to which social robots have recently been employed – storytelling, assisting human teachers, and language learning. There is no doubt that in each of those activities, sound design plays an important role. For example, in language learning, appropriate vocalizations and suitability of the robot's voice could improve the learners' experience and therefore catalyze the learning process. Belpaeme et al. (2018) also highlight the importance of appropriate sound design, listing it as one of the aspects hindering the wider expansion of robots in education. Therefore, there the development of social robots could help in achieving SDG 4 *Quality education* (Guenat et al., 2022; Nations, 2023). Furthermore, Guenat et al. (2022) underline the benefit of robots advancing gender equality by reducing the burden of low-paid tasks traditionally performed by women in agriculture, thereby freeing up time for education and contributing not only to SDG 4 but also SDG 5 – *Gender equality* (Guenat et al., 2022; Nations, 2023).

This replacement of repetitive tasks by social robots is also primarily the task of service robots (Mahdi et al., 2022). According to them such robots are also the most popular ones, right after research robots, and are employed in tasks such as cleaning or receptionist duties. Guenat et al. (2022) also emphasize the significance of RAS in substituting human tasks, particularly in areas experiencing labor shortages, which could contribute to meeting many different SDGs depending on the type of work and context in which the robot is placed.

While there are numerous benefits to the advancement of social robots, its crucial to acknowledge the potential threats that may arise from their development. Guenat et al. (2022) identify the most important threat as reinforcement of existing inequalities coming from the accessibility of technology to the wealthy. Moreover, they also highlight the transformation of the job market coming from the replacement of certain jobs by robots. Risks concerning negative environmental impacts, immense financial resources required for development and issues with governance, in particular with regards to data collection, also arise (Guenat et al., 2022). Furthermore, many risks coming from specific

implementations of robots can be defined. For example, as this project could contribute to anthropomorphizing robotic voices, the risk of confusing human-robot relationships in robots for elderly care can emerge (Nwosu et al., 2019). A broader example might be that students and teachers could struggle to accept and adapt to robots in educational settings, or that the robots in such settings might be inappropriately designed for the users' age groups. (Youssef et al., 2023).

Nonetheless, there is no doubt that the field of robotics will be developing. Therefore, gaining an understanding of the communication patterns between robots and humans, which also entails understanding the perception of robotic voices, can also help to understand the risks coming from certain robot designs and mitigate them.

References

101soundboards. (n.d.). Over 27,749,290 audio sound MP3 clips to play and download — 101soundboards.com [Accessed 16-05-2024].

American National Standards Institute. (1973). *American national standard psychoacoustical terminology*. American National Standards InstituteCommittee on Bioacoustics, S3; American National Standards Institute; Sonn, M.; Acoustical Society of America. https://books.google.nl/books?id=5Fo0GQAACAAJ

Arts, E. (2023). Ameca. *Engineered Arts*. https://www.engineeredarts.co.uk/robot/ameca/

ATLAS.ti version 24.1.0 for Mac. (2023). *Atlas.ti scientific software development gmbh [qualitative data analysis software]* (Version version 24.1.0 for Mac). https://atlasti.com

Avcontentteam. (2023). Recursive Feature Elimination: Working, Advantages & Examples — analyticsvidhya.com [Accessed 23-05-2024].

Bakardzhiev, H. (2022). *The role of voice character in navigational assistants: Prosodic differences and dialogue style's effect on perceptions of naturalness and anthropomorphism.* [Doctoral dissertation]. Retrieved May 1, 2024, from https://arno.uvt.nl/show.cgi?fid=157448

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40*, 531–539. https://doi.org/10.3758/brm.40.2.531

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*, *3*, eaat5954. https://doi.org/10.1126/scirobotics.aat5954

Bosland, R. (2022). *Automatic extraction of characterizing features for non-native dutch read speech* [Doctoral dissertation]. Retrieved May 14, 2024, from https://theses.ubn.ru.nl/server/api/core/bitstreams/dd89bd16-5981-41d3-96a6-6fa94ff1da67/content

Braun, V., & Clarke, V. (2012). Thematic analysis. *APA Handbook of Research Methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, *2*, 57–71. https://doi.org/10.1037/13620-004

Burke, J., Murphy, R., Rogers, E., Lumelsky, V., & Scholtz, J. (2004). Final report for the darpa/nsf interdisciplinary study on human-robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *34*(2), 103–112. https://doi.org/10.1109/TSMCC.2004.826287

Bäckström, T., Räsänen, O., Zewoudie, A., Zarazaga, P. P., Koivusalo, L., Das, S., Mellado, E. G., Mansali, M. B., Ramos, D., Kadiri, S., & Alku, P. (2022). Introduction to speech processing. https://doi.org/10.5281/zenodo.6821775

Cambridge University Press. (n.d.). *"robot"*. Retrieved April 28, 2024, from https://dictionary.cambridge.org/dictionary/english/robot

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*, 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Cifuentes, C. A., Pinto, M. J., Céspedes, N., & Múnera, M. (2020). Social robots in therapy and care. *Current Robotics Reports*, *1*, 59–74.

de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., & Wisniewski, G. Probing phoneme, language and speaker information in unsupervised speech representations. In: In *Proc. interspeech 2022*. 2022, 1402–1406. https://doi.org/10.21437/Interspeech.2022-373

Delić, V., Zoran, P., Sečujski, M., Jakovljevic, N., Nikolic, J., Miskovic, D., Simić, N., Suzic, S., & Delić, T. (2019). Speech technology progress based on new machine learning paradigm. *Computational Intelligence and Neuroscience*, *2019*, 1–19. https://doi.org/10.1155/2019/4368036

Dou, X., Wu, C.-F., Lin, K.-C., Gan, S., & Tseng, T.-M. (2020). Effects of different types of social robot voices on affective evaluations in different application fields. *International Journal of Social Robotics*, *13*, 615–628. https://doi.org/10.1007/s12369-020-00654-9

Duffy, B., Rooney, Duffy, B, Rooney, C, Oapos;hare, G, & Oapos;donoghue, R. (1999). Title what is a social robot? what is a social robot? https://researchrepository.ucd.ie/server/api/core/bitstreams/feefa084-a393-4368-8c72-b839c09c08b8/content

Ehret, J., Bönsch, A., Aspöck, L., Röhr, C. T., Baumann, S., Grice, M., Fels, J., & Kuhlen, T. W. (2021). Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? *ACM Transactions on Applied Perception*, *18*, 1–15. https://doi.org/10.1145/3486580

Ermers, L. (2023). *Ageing, perceived human-or robot-likeness and sound type: Exploring emotion recognition from semantic free utterances in hri* [Doctoral dissertation].

Retrieved May 3, 2024, from https://theses.ubn.ru.nl/server/api/core/bitstreams/7771eb1b-3f3f-430a-a2a7-a296ced62816/content

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, M., F.and Wöllmer, & Schuller, B. (2010). *Opensmile - the munich versatile and fast open-source audio feature extractor*. ACM Multimedia (MM), ACM, Florence, Italy.

Farnsworth, B. (2023). The science of voice analysis. https://imotions.com/blog/learning/research-fundamentals/the-science-of-voice-analysis/

Fosso, W. S., Queiroz, M. M., & Hamzi, L. (2023). A bibliometric and multi-disciplinary quasi-systematic analysis of social robots: Past, future, and insights of human-robot interaction. *Technological Forecasting and Social Change*, *197*, 122912. https://doi.org/10.1016/j.techfore.2023.122912

Gessinger, I., Cohn, M., Zellou, G., & Möbius, B. Cross-cultural comparison of gradient emotion perception: Human vs. alexa tts voices. In: In *Interspeech*. 2022, 4970–4974.

Grisel, O., Mueller, A., Lars, Gramfort, A., Louppe, G., Fan, T. J., Prettenhofer, P., Blondel, M., Niculae, V., Nothman, J., Lemaitre, G., Joly, A., Estève, L., du Boisberranger, J., Vanderplas, J., manoj kumar, Qin, H., Hug, N., Varoquaux, N., ... Marmo, C. (2024, May 13). *Scikit-learn/scikit-learn*. https://github.com/scikit-learn/scikit-learn

Guenat, S., Purnell, P., Davies, Z. G., Nawrath, M., Stringer, L. C., Babu, G. R., Balasubramanian, M., Ballantyne, E. E. F., Bylappa, B. K., Chen, B., De Jager, P., Del Prete, A., Di Nuovo, A., Ehi-Eromosele, C. O., Eskandari Torbaghan, M., Evans, K. L., Fraundorfer, M., Haouas, W., Izunobi, J. U., ... Dallimer, M. (2022). Meeting sustainable development goals via robotics and autonomous systems. *Nature Communications*, *13*. https://doi.org/10.1038/s41467-022-31150-5

Guizzo, E. (2018, August). Types of robots. *ROBOTS: Your Guide to the World of Robotics*. https://robotsguide.com/learn/types-of-robots/

Gurung, N., Grant, J. B., & Hearth, D. (2023). The uncanny effect of speech: The impact of appearance and speaking on impression formation in human–robot interactions. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-023-00976-4

Hansen, J., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *Signal Processing Magazine, IEEE*, *32*, 74–99. https://doi.org/10.1109/MSP.2015.2462851

Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2022). The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception. *Emotion (Washington, D.C.)*, *22*, 213–225. https://doi.org/10.1037/emo0001048

Jakubec, M., Jarina, R., Lieskovska, E., & Kasak, P. (2024). Deep speaker embeddings for speaker verification: Review and experimental comparison. *Engineering Applications of Artificial Intelligence*, *127*, 107232. https://doi.org/10.1016/j.engappai.2023.107232

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I., & Wu, Y. (2019). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Retrieved May 7, 2024, from https://arxiv.org/pdf/1806.04558

Kappas, A., Küster, D., Dente, P., & Basedow, C. (2014). Validated corpus of nonverbal acoustical emblems. *EMOTE*
Deliverable 3.2.

Kriz, S., Ferro, T. D., Damera, P., & Porter, J. R. (2010). Fictional robots as a data source in HRI research: Exploring the link between science fiction and interactional expectations, 458–463.

Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in Neurorobotics*, *14*. https://doi.org/10.3389/fnbot.2020.593732

Lam, D. (2023, November). Aramadia/willow-sound. *GitHub*. Retrieved May 3, 2024, from https://github.com/aramadia/willow-sound

Latupeirissa, A. B., Frid, E., & Bresin, R. (2019). Sonic characteristics of robots in films. *kth.diva-portal.org*. Retrieved May 2, 2024, from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-250788

Laukka, P., Elfenbein, H., Chui, W., Thingujam, N., Iraki, F., Rockstuhl, T., & Althoff, J. Presenting the venec corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In: 2010, January, 53–57.

Laver, J. (1994). The semiotic framework. In *Principles of phonetics* (13–25). Cambridge University Press.

Lee, J.-Y., Song, Y. A., Jung, J. Y., Kim, H. J., Kim, B. R., Do, H.-K., & Lim, J.-Y. (2018). Nurses' needs for care robots in integrated nursing care services. *Journal of Advanced Nursing*, *74*, 2094–2105. https://doi.org/10.1111/jan.13711

Lee, M. (2024, March). Best 6 robot voice changer in 2024. *EaseUS*. Retrieved May 3, 2024, from https://multimedia.easeus.com/voice-changer-tips/robot-voice-changer.html

Li, M., & Suh, A. Machinelike or humanlike? a literature review of anthropomorphism in ai-enabled technology. English. In: In *Proceedings of the 54th hawaii international*

conference on system sciences. Proceedings of the Annual Hawaii International Conference on System Sciences. 2021, January, 4053–4062. https://doi.org/10.24251/HICSS.2021.493

Lin, Y.-X., Cheng, C.-H., Le, P. T., Huang, B.-J., Chu-Xin, L., Huang, C.-L., & Wang, J.-C. (2023). Zero-shot voice conversion based on speaker embedding domain generalization. *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*. https://doi.org/10.1109/rivf60135.2023.10471830

Lu, H., Wu, X., Wu, Z., & Meng, H. Speechtriplenet: End-to-end disentangled speech representation learning for content, timbre and prosody. In: In *Proceedings of the 31st acm international conference on multimedia*. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, 2829–2837. ISBN: 9798400701085. https://doi.org/10.1145/3581783.3612485

Lüscher, C., Xu, J., Zeineldeen, M., Schlüter, R., & Ney, H. (2023, September). Analyzing and improving neural speaker embeddings for asr. *arXiv.org*. https://doi.org/10.48550/arXiv.2301.04571

Mahdi, H., Akgun, S. A., Saleh, S., & Dautenhahn, K. (2022). A survey on the design and evolution of social robots — past, present and future. *Robotics and Autonomous Systems*, *156*, 104193. https://doi.org/https://doi.org/10.1016/j.robot.2022.104193

Mara, M., Appel, M., & Gnambs, T. (2022). Human-like robots and the uncanny valley. *Zeitschrift für Psychologie*, *230*, 33–46. https://doi.org/10.1027/2151-2604/a000486

McGinn, C., & Torre, I. (2019). Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 211–221. https://doi.org/10.1109/hri.2019.8673305

Mende, M. A., Fischer, M. H., & Kühne, K. (2019). The use of social robots and the uncanny valley phenomenon. *AI Love You*, 41–73. https://doi.org/10.1007/978-3-030-19734-6_3

Microsoft/wavlm-large · hugging face. (n.d.). *huggingface.co*. Retrieved May 7, 2024, from https://huggingface.co/microsoft/wavlm-large

Moon, Y. (2001). Sony aibo: The world's first entertainment robot [5]02010. *HBR Store*. Retrieved April 30, 2024, from https://store.hbr.org/product/sony-aibo-the-world-s-first-entertainment-robot/502010?sku=502010-PDF-ENG

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, *7*, 33–35.

Naneva, S., Sarda Gou, M., Webb, T. L., & Prescott, T. J. (2020). A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *International Journal of Social Robotics*, *12*, 1179–1201. https://doi.org/10.1007/s12369-020-00659-4

Nations, U. (2023). United nations sustainable development. *United Nations*. https://www.un.org/sustainabledevelopment/

NCH Software. (n.d.). Voxal easy-to-use real-time voice changer software. download free. *www.nchsoftware.com*. https://www.nchsoftware.com/voicechanger/index.html

Network, V. (n.d.). Voicy: Free Sounds, Sound GIFs, SFX and Meme Soundboard — voicy.network [Accessed 16-05-2024].

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making social robots more attractive: The effects of voice pitch, humor and empathy. *International Journal of Social Robotics*, *5*, 171–191. https://doi.org/10.1007/s12369-012-0171-x

Nwosu, A. C., Sturgeon, B., McGlinchey, T., Goodwin, C. D., Behera, A., Mason, S., Stanley, S., & Payne, T. R. (2019). Robotic technology for palliative and supportive care: Strengths, weaknesses, opportunities and threats. *Palliative Medicine*, *33*, 1106–1113. https://doi.org/10.1177/0269216319857628

Nxbyte. (2024, April). Nxbyte/pythonaudioeffects. *GitHub*. Retrieved May 3, 2024, from https://github.com/nxbyte/PythonAudioEffects

Oliveira, R., & Yadollahi, E. (2024). Robots in movies: A content analysis of the portrayal of fictional social robots. *Behaviour & Information Technology*, *43*(5), 970–987. https://doi.org/10.1080/0144929X.2023.2196576

Oxford University Press. (2005). The oxford dictionary of phrase and fable (2nd ed.). Retrieved May 15, 2024, from https://www.oxfordreference.com/display/10.1093/acref/9780198609810.001.0001/acref-9780198609810?btog=chap&hide=true&jumpTo=robot&page=302&pageSize=20&skipEditions=true&sort=titlesort&source=%2F10.1093%2Facref%2F9780198609810.001.0001%2Facref-9780198609810
eISBN: 9780191727047.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qualtrics. (2005). Qualtrics. https://www.qualtrics.com
Copyright Year: 2024, Version: April, May 2024.

Rose, J. (2012). *Audio postproduction for film and video*. Taylor Francis.

Saleh, M. A., Hanapiah, F. A., & Hashim, H. (2020). Robot applications for autism: A comprehensive review. *Disability and Rehabilitation: Assistive Technology*, 1–23. https://doi.org/10.1080/17483107.2019.1685016

Savela, N., Turja, T., Latikka, R., & Oksanen, A. (2021). Media effects on the perceptions of robots. *Human Behavior and Emerging Technologies*, *3*. https://doi.org/10.1002/hbe2.296

Schreibelmayr, S., & Mara, M. (2022). Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.787499

Seabold, S., & Perktold, J. Statsmodels: Econometric and statistical modeling with python. In: In *9th python in science conference*. 2010.

Shaheen, Z., Sadekova, T., Matveeva, Y., Shirshova, A., & Kudinov, M. (2023). Exploiting emotion information in speaker embeddings for expressive text-to-speech. *INTERSPEECH 2023*. https://doi.org/10.21437/interspeech.2023-2407

Sheridan, T. B. (2020). A review of recent research in social robotics [Cyberpsychology]. *Current Opinion in Psychology*, *36*, 7–12. https://doi.org/https://doi.org/10.1016/j.copsyc.2020.01.003

Siedenburg, K., Saitis, C., & McAdams, S. (2019). The present, past, and future of timbre research. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, perception, and cognition* (pp. 1–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_1

Stan, A. (2022). Residual information in deep speaker embedding architectures. *Mathematics*, *10*, 3927. https://doi.org/10.3390/math10213927

Su, H., Qi, W., Chen, J., Yang, C., Sandoval, J., & Laribi, M. A. (2023). Recent advancements in multimodal human–robot interaction. *Front. Neurorobot.*, *17*. https://doi.org/10.3389/fnbot.2023.1084000

Tung, F.-W. (2016). Child perception of humanoid robot appearance and behavior. *International Journal of Human-Computer Interaction*, *32*, 493–502. https://doi.org/10.1080/10447318.2016.1172808

Ulgen, I. R., Du, Z., Busso, C., & Sisman, B. (2024, January). Revealing emotional clusters in speaker embeddings: A contrastive learning strategy for speech emotion recognition. *arXiv.org*. https://doi.org/10.48550/arXiv.2401.11017

Utsugi, A., Wang, H., & Ota, I. A voice quality analysis of japanese anime. In: 2019, August.

van der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-sne [Pagination: 27]. *Journal of Machine Learning Research*, *9*(nov), 2579–2605.

Vance, A. (2014, February). Willow garage's last days. *Bloomberg.com*. https://www.bloomberg.com/news/articles/2014-02-20/robotics-research-lab-willow-garage-shuts-down

Voicemod. (2019). Free real time voice changer and modulator - voicemod. *Voicemod - Real Time Voice Changer Technology*. https://www.voicemod.net/

VoiceWave. (n.d.). Transform your voice in real-time for free with easeus voicewave! *EaseUS® Multimedia — Your Must-Have Video Audio Toolkit*. Retrieved May 3, 2024, from https://multimedia.easeus.com/voice-changer/

Wilson, S., & Moore, R. (2017). Robot, alien and cartoon voices: Implications for speech-enabled systems. https://vihar-2017.vihar.org/assets/papers/VIHAR-2017_paper_1.pdf

Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E., & Lefeber, D. (2011). Emogib: Emotional gibberish speech database for affective human-robot interaction. *Affective Computing and Intelligent Interaction*, 163–172. https://doi.org/10.1007/978-3-642-24571-8_17

Yilmazyildiz, S., Read, R., Belpeame, T., & Verhelst, W. (2015). Review of semantic free utterances in social human-robot interaction. *International Journal of Human-Computer Interaction*, *32*. https://doi.org/10.1080/10447318.2015.1093856

Youssef, K., Said, S., Alkork, S., & Beyrouthy, T. (2023). Social robotics in education: A survey on recent studies and applications. *International Journal of Emerging Technologies in Learning (Online)*, *18*(3), 67.

Yu, C., Fu, C., Chen, R., & Tapus, A. (2022). First attempt of gender-free speech style transfer for genderless robot. *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. https://doi.org/10.1109/hri53351.2022.9889533

Appendix

A eGeMAPS Features Analysis

Table 5: Results of linear regression analysis per each eGeMAPS feature and *roboticess*

	R^2	MSE	Slope	Intercept
slopeUV0-500_sma3nz_amean	0.299335	2.249665	17.137698	4.250508
loudness_sma3_amean	0.181151	2.629124	-0.853551	5.965060
loudness_sma3_percentile80.0	0.168286	2.670431	-0.568994	6.174027
loudness_sma3_percentile20.0	0.124348	2.811506	-0.859985	4.977653
VoicedSegmentsPerSec	0.110824	2.854928	-0.510426	5.592328
slopeUV500-1500_sma3nz_amean	0.109528	2.859088	52.252516	4.354864
equivalentSoundLevel_dBp	0.104371	2.875646	-0.101406	2.558274
loudness_sma3_percentile50.0	0.088680	2.926027	-0.478663	5.293782
MeanUnvoicedSegmentLength	0.086170	2.934084	2.279525	4.248957
loudness_sma3_stddevRisingSlope	0.082099	2.947156	-0.076197	5.318106
spectralFlux_sma3_amean	0.060683	3.015918	-0.500864	5.281241
loudness_sma3_pctlrange0-2	0.058471	3.023019	-0.386798	5.543520
F1amplitudeLogRelF0_sma3nz_amean	0.054361	3.036217	-0.008035	3.988153
MeanVoicedSegmentLengthSec	0.052961	3.040711	0.521741	4.339675

Continued on next page

Feature	R^2	MSE	Slope	Intercept
loudness_sma3_meanRisingSlope	0.052714	3.041504	-0.032146	5.189713
loudness_sma3_stddevFallingSlope	0.048689	3.054428	-0.071424	5.065387
hammarbergIndexV_sma3nz_amean	0.045739	3.063899	0.036470	3.949524
F3bandwidth_sma3nz_stddevNorm	0.043712	3.070408	-1.967366	5.508401
F1frequency_sma3nz_stddevNorm	0.041993	3.075927	-2.969483	5.507104
loudness_sma3_stddevNorm	0.040336	3.081247	1.061465	3.697821
F3amplitudeLogRelF0_sma3nz_amean	0.038446	3.087315	-0.008654	3.999021
F2amplitudeLogRelF0_sma3nz_amean	0.037212	3.091276	-0.008047	4.081111
F3bandwidth_sma3nz_amean	0.034333	3.100521	0.001456	3.477674
spectralFlux_sma3_stddevNorm	0.032214	3.107323	0.868751	3.679612
F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	0.028176	3.120288	-0.001615	4.800533
F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope	0.027727	3.121731	-0.001674	4.849142
spectralFluxV_sma3nz_amean	0.025749	3.128081	-0.235998	5.091296
loudness_sma3_meanFallingSlope	0.023772	3.134429	-0.023766	4.946173
logRelF0-H1-H2_sma3nz_amean	0.019550	3.147985	0.019602	4.536781
loudnessPeaksPerSec	0.019226	3.149025	-0.175969	5.096621
alphaRatioUV_sma3nz_amean	0.018738	3.150591	0.031876	4.787154
F2frequency_sma3nz_stddevNorm	0.016419	3.158039	-4.383039	5.239989
logRelF0-H1-A3_sma3nz_amean	0.015280	3.161695	0.017143	4.419625
mfcc2_sma3_amean	0.014775	3.163315	0.015466	4.840077
mfcc4V_sma3nz_amean	0.013960	3.165932	0.013264	4.879019
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	0.013546	3.167263	0.021165	3.688478
F2frequency_sma3nz_amean	0.012460	3.170751	0.000832	3.181113
F3frequency_sma3nz_amean	0.011664	3.173306	0.000672	2.747029
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	0.010989	3.175471	0.029002	4.433288
F2bandwidth_sma3nz_stddevNorm	0.010707	3.176378	-0.917400	4.981316
F1bandwidth_sma3nz_stddevNorm	0.009658	3.179745	1.306727	4.341989
logRelF0-H1-H2_sma3nz_stddevNorm	0.009064	3.181654	0.010429	4.648675
F1frequency_sma3nz_amean	0.008319	3.184044	0.000797	4.013329
mfcc2V_sma3nz_amean	0.008125	3.184669	0.009853	4.836711
spectralFluxUV_sma3nz_amean	0.007910	3.185360	-0.379154	4.738594
F0semitoneFrom27.5Hz_sma3nz_amean	0.007474	3.186758	0.015620	4.003048
F1bandwidth_sma3nz_amean	0.007249	3.187481	-0.000642	5.379172
mfcc1V_sma3nz_stddevNorm	0.005967	3.191598	-0.028562	4.663001
F1amplitudeLogRelF0_sma3nz_stddevNorm	0.005637	3.192657	-0.059679	4.588946
StddevUnvoicedSegmentLength	0.005626	3.192692	0.956249	4.584253
slopeV0-500_sma3nz_amean	0.005249	3.193901	-3.183134	4.850738

Continued on next page

Feature	R^2	MSE	Slope	Intercept
shimmerLocaldB_sma3nz_stddevNorm	0.005156	3.194201	-0.305494	4.889193
mfcc2_sma3_stddevNorm	0.004717	3.195610	0.001612	4.662435
alphaRatioV_sma3nz_stddevNorm	0.004546	3.196160	-0.000271	4.661524
HNRdBACF_sma3nz_amean	0.004373	3.196714	0.021909	4.497885
mfcc1_sma3_amean	0.004348	3.196796	-0.009746	4.772759
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	0.004238	3.197148	0.011484	4.171962
logRelF0-H1-A3_sma3nz_stddevNorm	0.004135	3.197479	-0.005929	4.654748
mfcc3V_sma3nz_stddevNorm	0.003170	3.200578	0.001048	4.657065
spectralFluxV_sma3nz_stddevNorm	0.003130	3.200706	-0.447513	4.969482
F2bandwidth_sma3nz_amean	0.002664	3.202203	0.000399	4.281080
mfcc1V_sma3nz_amean	0.002661	3.202211	0.005930	4.562930
mfcc3_sma3_amean	0.002044	3.204191	0.006301	4.694919
mfcc4V_sma3nz_stddevNorm	0.001977	3.204409	-0.000705	4.659397
mfcc4_sma3_amean	0.001759	3.205107	0.006508	4.716428
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	0.001734	3.205187	0.006642	4.403681
mfcc2V_sma3nz_stddevNorm	0.001582	3.205675	-0.003054	4.658547
F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	0.001510	3.205908	-0.000924	4.700604
F2amplitudeLogRelF0_sma3nz_stddevNorm	0.001116	3.207171	0.003016	4.661537
mfcc3V_sma3nz_amean	0.001095	3.207239	0.003904	4.693980
hammarbergIndexUV_sma3nz_amean	0.000944	3.207723	0.005898	4.579172
F3amplitudeLogRelF0_sma3nz_stddevNorm	0.000912	3.207828	-0.020421	4.630589
HNRdBACF_sma3nz_stddevNorm	0.000880	3.207931	-0.009293	4.659522
hammarbergIndexV_sma3nz_stddevNorm	0.000811	3.208151	-0.068494	4.681246
shimmerLocaldB_sma3nz_amean	0.000682	3.208565	-0.082355	4.749724
mfcc1_sma3_stddevNorm	0.000582	3.208887	0.004257	4.653289
slopeV0-500_sma3nz_stddevNorm	0.000462	3.209272	-0.003271	4.661344
mfcc4_sma3_stddevNorm	0.000454	3.209298	-0.001660	4.658018
alphaRatioV_sma3nz_amean	0.000420	3.209408	-0.002995	4.641502
jitterLocal_sma3nz_stddevNorm	0.000407	3.209447	-0.060309	4.733504
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	0.000379	3.209538	-0.000370	4.680156
jitterLocal_sma3nz_amean	0.000315	3.209744	1.141732	4.617456
slopeV500-1500_sma3nz_amean	0.000285	3.209840	-1.797536	4.644242
mfcc3_sma3_stddevNorm	0.000164	3.210228	-0.000075	4.657431
F3frequency_sma3nz_stddevNorm	0.000078	3.210504	0.470301	4.613630
slopeV500-1500_sma3nz_stddevNorm	0.000055	3.210579	0.003939	4.658935
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	0.000014	3.210710	0.069237	4.647839
StddevVoicedSegmentLengthSec	0.000006	3.210737	-0.019879	4.658546

Table 6: Results of linear regression analysis per each eGeMAPS feature and suitability for social robots

	R^2	MSE	Slope	Intercept
slopeUV0-500_sma3nz_amean	0.236351	0.970583	-9.581183	3.769536
slopeUV500-1500_sma3nz_amean	0.089377	1.157384	-29.697822	3.713991
VoicedSegmentsPerSec	0.066660	1.186257	0.249066	3.085925
loudness_sma3_stddevRisingSlope	0.065339	1.187936	0.042769	3.171206
loudness_sma3_percentile20.0	0.047125	1.211087	0.333090	3.418199
loudness_sma3_meanRisingSlope	0.042968	1.216370	0.018260	3.239672
MeanVoicedSegmentLengthSec	0.040381	1.219657	-0.286638	3.716619
loudness_sma3_amean	0.038123	1.222527	0.246360	3.164917
equivalentSoundLevel_dBp	0.036336	1.224798	0.037645	4.321523
loudness_sma3_stddevFallingSlope	0.035203	1.226239	0.038211	3.323814
MeanUnvoicedSegmentLength	0.031725	1.230660	-0.870221	3.698171
loudness_sma3_percentile80.0	0.029665	1.233277	0.150305	3.141740
mfcc4V_sma3nz_amean	0.028690	1.234516	-0.011964	3.341794
alphaRatioUV_sma3nz_amean	0.027331	1.236244	-0.024221	3.443243
alphaRatioV_sma3nz_amean	0.025689	1.238331	-0.014745	3.469701
mfcc1_sma3_amean	0.024965	1.239251	0.014694	3.367108
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	0.022384	1.242532	-0.017118	4.325433
shimmerLocaldB_sma3nz_stddevNorm	0.021621	1.243501	0.393602	3.242623
F1bandwidth_sma3nz_amean	0.019723	1.245914	0.000666	2.792488
loudness_sma3_meanFallingSlope	0.019699	1.245944	0.013612	3.376647
loudnessPeaksPerSec	0.018638	1.247292	0.109008	3.269904
slopeV500-1500_sma3nz_amean	0.017586	1.248630	-8.884087	3.482951
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	0.015259	1.251587	-0.013710	4.120902
mfcc1V_sma3nz_amean	0.014901	1.252042	0.008829	3.403614
F0semitoneFrom27.5Hz_sma3nz_amean	0.013950	1.253252	-0.013426	4.104172
spectralFluxV_sma3nz_stddevNorm	0.013078	1.254359	0.575539	3.139907
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	0.012648	1.254906	-0.019576	3.693202
F1frequency_sma3nz_amean	0.012448	1.255160	-0.000614	4.037514
F2frequency_sma3nz_amean	0.011586	1.256255	-0.000505	4.437698
mfcc1V_sma3nz_stddevNorm	0.010806	1.257247	0.024184	3.537002
F3bandwidth_sma3nz_stddevNorm	0.010407	1.257755	0.603956	3.281079
loudness_sma3_percentile50.0	0.010065	1.258188	0.101460	3.407537
logRelF0-H1-H2_sma3nz_amean	0.008877	1.259699	-0.008310	3.593337
spectralFlux_sma3_amean	0.008500	1.260178	0.117940	3.395505
slopeV0-500_sma3nz_amean	0.008410	1.260292	2.534989	3.387828

Continued on next page

Feature	R^2	MSE	Slope	Intercept
F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope	0.008236	1.260513	0.000574	3.476537
mfcc2_sma3_stddevNorm	0.008048	1.260752	-0.001325	3.537634
F1bandwidth_sma3nz_stddevNorm	0.007695	1.261201	-0.733843	3.719183
shimmerLocaldB_sma3nz_amean	0.007642	1.261268	-0.173431	3.739350
jitterLocal_sma3nz_stddevNorm	0.007231	1.261791	0.159846	3.338091
F3amplitudeLogRelF0_sma3nz_stddevNorm	0.007151	1.261892	0.035987	3.588004
logRelF0-H1-A3_sma3nz_stddevNorm	0.006964	1.262130	0.004841	3.543942
F1amplitudeLogRelF0_sma3nz_amean	0.006756	1.262394	0.001782	3.690862
F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	0.006633	1.262551	0.000493	3.498635
mfcc4_sma3_amean	0.006132	1.263187	-0.007644	3.472055
F1frequency_sma3nz_stddevNorm	0.005998	1.263357	0.706108	3.340351
F3bandwidth_sma3nz_amean	0.005483	1.264012	-0.000366	3.839020
mfcc2V_sma3nz_amean	0.005125	1.264468	0.004923	3.632794
F2bandwidth_sma3nz_stddevNorm	0.004759	1.264932	0.384821	3.406332
mfcc2_sma3_amean	0.004438	1.265341	0.005333	3.606018
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	0.004407	1.265380	-0.006661	3.796046
F3frequency_sma3nz_amean	0.003707	1.266269	-0.000238	4.219886
jitterLocal_sma3nz_amean	0.003583	1.266427	-2.422370	3.625119
spectralFluxV_sma3nz_amean	0.003577	1.266434	0.055345	3.440649
loudness_sma3_pctlrange0-2	0.003164	1.266960	0.056610	3.412810
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	0.002765	1.267466	-0.000628	3.583149
F2frequency_sma3nz_stddevNorm	0.002505	1.267798	1.077084	3.399227
logRelF0-H1-H2_sma3nz_stddevNorm	0.002463	1.267850	-0.003421	3.545165
mfcc3V_sma3nz_amean	0.002409	1.267919	0.003644	3.577798
hammarbergIndexV_sma3nz_stddevNorm	0.002294	1.268065	0.072472	3.516286
F1amplitudeLogRelF0_sma3nz_stddevNorm	0.002287	1.268074	0.023919	3.569661
mfcc3_sma3_amean	0.002252	1.268119	0.004161	3.568143
mfcc1_sma3_stddevNorm	0.002028	1.268403	-0.005001	3.546220
slopeV500-1500_sma3nz_stddevNorm	0.001924	1.268536	-0.014690	3.532914
F2bandwidth_sma3nz_amean	0.001859	1.268618	-0.000210	3.739886
spectralFluxUV_sma3nz_amean	0.001823	1.268664	0.114528	3.517805
HNRdBACF_sma3nz_amean	0.001752	1.268755	0.008724	3.479568
F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	0.001487	1.269091	-0.000577	3.570308
loudness_sma3_stddevNorm	0.001252	1.269389	-0.117680	3.648921
hammarbergIndexV_sma3nz_amean	0.000937	1.269790	-0.003285	3.606318
mfcc2V_sma3nz_stddevNorm	0.000822	1.269937	0.001385	3.541648
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	0.000784	1.269984	-0.325614	3.582556

Continued on next page

Feature	R^2	MSE	Slope	Intercept
F3frequency_sma3nz_stddevNorm	0.000751	1.270026	-0.917772	3.625970
HNRdBACF_sma3nz_stddevNorm	0.000706	1.270084	-0.005236	3.544459
StddevUnvoicedSegmentLength	0.000589	1.270232	-0.194709	3.557331
alphaRatioV_sma3nz_stddevNorm	0.000504	1.270341	-0.000057	3.543746
mfcc4_sma3_stddevNorm	0.000459	1.270397	0.001051	3.541584
mfcc4V_sma3nz_stddevNorm	0.000338	1.270551	0.000183	3.541856
F2amplitudeLogRelF0_sma3nz_amean	0.000315	1.270580	0.000466	3.575965
mfcc3V_sma3nz_stddevNorm	0.000273	1.270635	-0.000193	3.542520
F2amplitudeLogRelF0_sma3nz_stddevNorm	0.000171	1.270763	-0.000744	3.541371
F3amplitudeLogRelF0_sma3nz_amean	0.000148	1.270793	0.000338	3.568343
StddevVoicedSegmentLengthSec	0.000113	1.270837	-0.055981	3.548919
spectralFlux_sma3_stddevNorm	0.000083	1.270876	-0.027745	3.573850
slopeV0-500_sma3nz_stddevNorm	0.000082	1.270876	0.000869	3.541322
mfcc3_sma3_stddevNorm	0.000074	1.270887	0.000032	3.542188
hammarbergIndexUV_sma3nz_amean	0.000065	1.270898	-0.000977	3.555433
logRelF0-H1-A3_sma3nz_amean	0.000034	1.270938	0.000511	3.535603

A.1 RFE Statistics

Table 7: RFE results per number of features selected

No. Features	R^2 *roboticness*	MSE *roboticness*	R^2 suitability	MSE suitability
1	0.046969	-3.042429	0.021466	-1.235642
2	0.057478	-3.005977	0.030999	-1.223322
3	0.044520	-3.046881	0.021343	-1.235905
4	0.165504	-2.668554	0.036577	-1.216098
5	0.182476	-2.615267	0.037519	-1.214568
6	0.193393	-2.580785	0.034119	-1.219065
7	0.359629	-2.066562	0.252137	-0.944585
8	0.369741	-2.033151	0.269575	-0.922549
9	0.419138	-1.856005	0.254530	-0.940786
10	0.414677	-1.870867	0.248449	-0.948704
11	0.436134	-1.799528	0.208034	-1.002963
12	0.418972	-1.850317	0.185329	-1.034886
13	0.429644	-1.817046	0.184384	-1.036226
14	0.436476	-1.796063	0.181549	-1.039884

Continued on next page

No. Features	R^2 *roboticness*	MSE *roboticness*	R^2 suitability	MSE suitability
15	0.439263	-1.788577	0.189762	-1.029675
16	0.434623	-1.801698	0.177305	-1.046629
17	0.418259	-1.849137	0.183604	-1.037809
18	0.420740	-1.842373	0.195419	-1.023419
19	0.416599	-1.854630	0.198965	-1.018813
20	0.427362	-1.821031	0.205085	-1.011047
21	0.428097	-1.818697	0.210979	-1.004156
22	0.440053	-1.781343	0.206652	-1.009903
23	0.454094	-1.737825	0.219292	-0.990882
24	0.457719	-1.727778	0.208291	-1.005979
25	0.455541	-1.731971	0.208079	-1.005648
26	0.463745	-1.703878	0.229402	-0.978943
27	0.464427	-1.702007	0.232171	-0.975736
28	0.456245	-1.729500	0.269235	-0.926175
29	0.464819	-1.705926	0.272581	-0.922224
30	0.467958	-1.696158	0.273993	-0.919712
31	0.472239	-1.683913	0.285554	-0.906078
32	0.474782	-1.677013	0.274045	-0.920110
33	0.489409	-1.629520	0.261267	-0.937157
34	0.489085	-1.630427	0.267092	-0.929233
35	0.493684	-1.614991	0.264396	-0.932416
36	0.478833	-1.663468	0.258240	-0.939439
37	0.468939	-1.693005	0.264315	-0.931104
38	0.466829	-1.699215	0.273192	-0.921243
39	0.461096	-1.717074	0.287985	-0.902900
40	0.457242	-1.730105	0.279780	-0.913919
41	0.463519	-1.711153	0.276654	-0.917855
42	0.458710	-1.726163	0.277871	-0.914681
43	0.461329	-1.719042	0.279548	-0.912067
44	0.458907	-1.726188	0.280822	-0.910432
45	0.452878	-1.744382	0.282332	-0.908518
46	0.451320	-1.751616	0.284679	-0.905260
47	0.453511	-1.743841	0.282355	-0.908185
48	0.451831	-1.749345	-14.046248	-20.496742
49	0.456764	-1.732737	-14.267669	-20.795730
50	0.461340	-1.719952	-14.440849	-21.031691
51	0.467414	-1.700972	-14.437436	-21.027163

Continued on next page

No. Features	R^2 *roboticness*	MSE *roboticness*	R^2 suitability	MSE suitability
52	-11.253520	-37.330176	-19.936653	-28.594001
53	-11.307473	-37.493906	-16.832512	-24.349937
54	-12.081153	-39.843911	-19.277236	-27.712309
55	-12.507967	-41.148793	-18.686716	-26.899708
56	-14.561244	-47.389926	-18.711095	-26.933508
57	-14.007287	-45.710849	-18.272061	-26.328255
58	-14.587240	-47.484416	-17.933867	-25.861177
59	-13.834825	-45.196350	-18.843149	-27.048207
60	-14.359875	-46.792314	-19.118078	-27.427195
61	-14.358192	-46.788584	-18.944210	-27.190277
62	-13.803475	-45.101438	-18.914147	-27.142949
63	-13.996535	-45.688135	-19.061912	-27.347582
64	-8.328596	-28.503968	-19.077708	-27.368293
65	-8.252981	-28.273875	-19.269490	-27.631430
66	-7.592015	-26.262668	-19.472100	-27.880941
67	-7.563259	-26.178223	-21.518052	-30.694302
68	-7.553037	-26.145897	-19.986302	-28.573766
69	-7.536183	-26.099369	-20.051272	-28.650874
70	-8.794826	-29.925311	-22.542798	-32.083598
71	-9.018874	-30.597611	-21.297341	-30.359362
72	-8.378287	-28.646285	-21.334826	-30.424409
73	-8.435224	-28.824893	-21.098287	-30.080621
74	-8.235761	-28.219201	-20.691348	-29.528345
75	-8.230942	-28.202046	-20.521746	-29.292806
76	-8.481881	-28.953968	-20.906972	-29.812511
77	-8.295332	-28.383997	-20.777516	-29.646318
78	-8.123827	-27.862668	-20.278287	-28.959662
79	-8.618290	-29.368742	-20.483328	-29.229956
80	-8.267255	-28.307429	-20.466549	-29.207813
81	-7.607393	-26.301152	-20.453002	-29.189123
82	-7.447407	-25.812099	-20.005357	-28.571820
83	-7.445815	-25.806911	-20.345507	-29.036626
84	-7.447312	-25.807340	-20.298506	-28.973745
85	-7.406409	-25.683508	-20.353364	-29.049037
86	-7.083903	-24.702805	-20.375410	-29.080135
87	-6.945976	-24.284114	-20.463426	-29.201296
88	-6.949811	-24.296028	-20.487312	-29.234345

A.2 35 Best Features Selected by RFE

Table 8: 35 best features predicting *roboticness*

Feature Name
F0semitoneFrom27.5Hz_sma3nz_stddevNorm
F0semitoneFrom27.5Hz_sma3nz_percentile20.0
F0semitoneFrom27.5Hz_sma3nz_percentile80.0
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2
loudness_sma3_amean
loudness_sma3_percentile20.0
loudness_sma3_percentile50.0
loudness_sma3_percentile80.0
loudness_sma3_pctlrange0-2
spectralFlux_sma3_amean
spectralFlux_sma3_stddevNorm
mfcc4_sma3_amean
jitterLocal_sma3nz_amean
shimmerLocaldB_sma3nz_amean
shimmerLocaldB_sma3nz_stddevNorm
F1frequency_sma3nz_stddevNorm
F1bandwidth_sma3nz_stddevNorm
F2frequency_sma3nz_stddevNorm
F2bandwidth_sma3nz_stddevNorm
F3frequency_sma3nz_stddevNorm
F3bandwidth_sma3nz_stddevNorm
hammarbergIndexV_sma3nz_stddevNorm
slopeV0-500_sma3nz_amean
slopeV500-1500_sma3nz_amean
spectralFluxV_sma3nz_amean
spectralFluxV_sma3nz_stddevNorm
slopeUV0-500_sma3nz_amean
slopeUV500-1500_sma3nz_amean
spectralFluxUV_sma3nz_amean
loudnessPeaksPerSec
VoicedSegmentsPerSec
MeanVoicedSegmentLengthSec
StddevVoicedSegmentLengthSec

Continued on next page

Feature Name
MeanUnvoicedSegmentLength
StddevUnvoicedSegmentLength

Table 9: 39 best features predicting suitability for social robots

Feature Name
F0semitoneFrom27.5Hz_sma3nz_stddevNorm
F0semitoneFrom27.5Hz_sma3nz_percentile20.0
F0semitoneFrom27.5Hz_sma3nz_percentile50.0
F0semitoneFrom27.5Hz_sma3nz_percentile80.0
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2
loudness_sma3_amean
loudness_sma3_stddevNorm
loudness_sma3_percentile20.0
loudness_sma3_percentile50.0
loudness_sma3_percentile80.0
loudness_sma3_pctlrange0-2
spectralFlux_sma3_amean
spectralFlux_sma3_stddevNorm
mfcc4_sma3_amean
jitterLocal_sma3nz_amean
shimmerLocaldB_sma3nz_amean
shimmerLocaldB_sma3nz_stddevNorm
F1frequency_sma3nz_stddevNorm
F1bandwidth_sma3nz_stddevNorm
F2frequency_sma3nz_stddevNorm
F2bandwidth_sma3nz_stddevNorm
F3frequency_sma3nz_stddevNorm
F3bandwidth_sma3nz_stddevNorm
hammarbergIndexV_sma3nz_stddevNorm
slopeV0-500_sma3nz_amean
slopeV500-1500_sma3nz_amean
spectralFluxV_sma3nz_amean
spectralFluxV_sma3nz_stddevNorm
mfcc4V_sma3nz_amean

Continued on next page

Feature Name
slopeUV0-500_sma3nz_amean
slopeUV500-1500_sma3nz_amean
spectralFluxUV_sma3nz_amean
loudnessPeaksPerSec
VoicedSegmentsPerSec
MeanVoicedSegmentLengthSec
StddevVoicedSegmentLengthSec
MeanUnvoicedSegmentLength
StddevUnvoicedSegmentLength
equivalentSoundLevel_dBp